DOCUMENT RESUME

ED 222 555 TM 820 709

AUTHOR	Folsom, Ralph E.; Williams, Rick L.
TITLE	Design Effects and the Analysis of Survey Data.
INSTITUTION	Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.; Research Triangle Inst., Research Triangle Park, N.C.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
REPORT NO	RTI/2137-01-01F
PUB DATE	Apr 82
GRANT	NIE-G-80-0003
NOTE	153p.; Appendix B is marginally legible due to small print; For related documents, see TM 820 707-712 and TM 820 716.
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Cluster Analysis; Educational Assessment; *Hypothesis Testing; *Mathematics Achievement; National Surveys; Probability; *Research Design; *Sampling; Secondary Education; *Statistical Analysis
IDENTIFIERS	Chi Square; National Assessment of Educational Progress; *NIE ECS NAEP Item Development Project; Second Mathematics Assessment (1978)

ABSTRACT

The National Assessment of Educational Progress (NAEP), like most large national surveys, employs a complex stratified multistage unequal probability sample. The design provides a rigorous justification for extending survey results to the entire U.S. target population. Developments in the analysis of data from complex surveys which provide a straightforward method for taking account of the sample design through proper estimation of subpopulation estimates and their covariance matrix are reviewed. Relationships among subpopulations are then evaluated via large sample Wald statistics assumed to be asymptotically distributed as central chi-squared random variables. While these methods provide a mechanism for analyzing NAEP data, the computer software required to properly estimate sample design-based covariance matrices is not generally available to NAEP data users. Design effect methods for adjusting test statistics obtained from standard statistical methods which implicitly assume simple random sampling from an infinite population are presented with several new decompositions obtained which display the effects of multistage clustering, stratification, and unequal weighting on the covariance matrix. A comparison of asymptotically valid sample design-based chi-squared tests versus analogous simple random sampling tests and design effect adjusted tests is given. Design effect adjustments on NAEP mathematics data are shown. Primary type of information provided by the report: Procedures (Sampling); Results (Secondary Analysis). (Author/CM)

****
*   Reproductions supplied by EDRS are the best that can be made   *
*   from the original document.   *
****

ED222555

RTI/2137/01-01F

DESIGN EFFECTS AND THE ANALYSIS OF SURVEY DATA

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- ☒ This document has been reproduced as received from the person or organization originating it.
  Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

by

Ralph E. Folsom
and
Rick L. Williams

Prepared for

Education Commission of the States
Denver, Colorado

"The work upon which this publication is based is performed pursuant to Grant NIE-G-80-0003 of the National Institute of Education. It does not, however, necessarily reflect the view of that agency."

April 1982

TM 820 709

RESEARCH TRIANGLE PARK, NORTH CAROLINA 27709

Abstract

Design Effects and the Analysis of Survey Data

by
Ralph E. Folsom
and
Rick L. Williams

The National Assessment of Educational Progress (NAEP), like most large national surveys, employs a complex stratified multistage unequal probability sample. When properly accounted for in the analysis, the NAEP sample design provides a rigorous justification for extending survey results to the entire U.S. student target population. This paper reviews recent developments in the analysis of data from complex surveys which provide a straightforward method for taking account of the sample design through proper estimation of subpopulation estimates and their covariance matrix. Relationships among subpopulations can then be evaluated via large sample Wald statistics assumed to be asymptotically distributed as central chi-squared random variables.

While these methods provide a mechanism for analyzing NAEP data, the computer software required to properly estimate sample design-based covariance matrices is not generally available to NAEP data users. Recent literature has suggested methods for adjusting test statistics obtained from standard statistical methods which implicitly assume simple random sampling from an infinite population. These so called design effect adjustments are reviewed and several new decompositions obtained which display the effects of multistage clustering, stratification and unequal weighting on the covariance matrix.

Finally, an empirical comparison is presented of asymptotically valid sample design-based chi-squared tests versus analogous simple random sampling tests and design effect adjusted tests. These comparisons are made for linear contrasts of domain means and proportions as well as for linear models fitted to the domain estimates via weighted least squares. The data were taken from the NAEP 1977-78 Mathematics assessment for 9-, 13- and 17-year-olds. For these data, the analyses indicate that the design effect type adjustments of standard test statistics are not stable and are generally too conservative to be of practical value.

TABLE OF CONTENTS

Page

1. INTRODUCTION 1

2. THEORETICAL RESULTS 3

2.1 The Wald Statistic/Weighted Least-Square Theory . . . 3
2.2 Generalized Design Effect Methods 7
2.3 Design Effect Models for P-Value Covariance Matrices 13

2.3.1 Single Stage Covariance Matrix Models 16
2.3.2 Three-Stage Covariance Matrix Models for Estimated Totals 19
2.3.3 Design Effect Model for the NAEP P-Value Covariance Matrix 28
2.3.4 Estimates for Composite Covariance Matrix Components 35
2.3.5 Asymptotic Distribution of SRS based NAEP Wald Statistics 42

2.4 Testing Balanced Fits Via Dummy Variable Regression . 48
2.5 Inference for NAEP Package Means 54

3. EMPIRICAL INVESTIGATION 60

3.1 Analysis Items and Subgroups 60
3.2 Analyses 60
3.3 Results 65

3.3.1 Wald Statistic/Weight Least Squares 65
3.3.2 Balanced Effects 84

4. COMMENTS ON NAEP DATA AND DOCUMENTATION 86

REFERENCES. 87
Appendix A: NAEP Exercises A-1
Appendix B: Wald Statistic Chi Squareds. B-1
Appendix C: Balanced Effect F-Tests. C-1
Appendix D: Contingency Tables of Wald Statistic Sample Design Based Tests Verses Alternative Tests Accepted and Rejected at the 5% Significance Level. D-1
Appendix E: Contingency Tables of Balanced Effects Sample Design Based Tests Versus Alternative Tests Accepted and Rejected at the 5% Significance Level. E-1

LIST OF TABLES

Table

3-1 Hypothesis Tests for the Race*Sex*PARED Cross-Classification 63

3-2 Hypothesis Tests for the Sex*TOC*PARED Cross-Classification 64

3-3 NAEP Item Design Effects for the Race*Sex*PARED Cross-Classification 67

3-4 NAEP Item Design Effects for the Sex*TOC*PARED Cross-Classification 68

3-5 Mean Scores Design Effects 69

3-6 Hypothesis Test Design Effects by NAEP Item for the Race*Sex*PARED Cross-Classification 71

3-7 Hypothesis Test Design Effects by NAEP Item for the Sex*TOC*PARED Cross-Classification 72

3-8 Hypothesis Test Design Effects for Mean Scores . . . 73

3-9 Hypothesis Test Design Effects for NAEP Items by Test Number for the Race*Sex*PARED Cross-Classification 75

3-10 Hypothesis Test Design Effects for NAEP Items by Test Number for the Sex*TOC*PARED Cross-Classification . . 76

3-11 Hypothesis Test Design Effects for Mean Scores by Test Number for the Race*Sex*PARED Cross-Classification 77

3-12 Hypothesis Test Design Effects for Mean Scores by Test Number for the Sex*TOC*PARED Cross-Classification 78

3-13 Conditional Percent of Contrast Design Based (DB) Tests Versus Alternative Tests (AT) Reaching an Opposite Conclusion for NAEP Items 80

3-14 Conditional Percent of Contrast Design Based (DB) Test Versus Alternative Tests (AT) Reaching an Opposite Conclusion for Mean Scores 81

3-15 Conditional Percent of Linear Model Design Based (DB) Tests Versus Alternative Tests (AT) Reaching an Opposite Conclusion for NAEP Items 82

LIST OF TABLES (continued)

Table		Page
3-16	Conditional Percent of Linear Model Design Based (DB) Tests Versus Alternative Tests (AT) Reaching an Opposite Conclusion for Mean Scores	83
3-17	Conditional Percent of Balanced Effects Design Based (DB) Test Versus Alternative Tests (AT) Reaching an Opposite Conclusion	85

6

1. INTRODUCTION

The National Assessment data base provides a wealth of information on the way student ability to correctly answer NAEP test items relates to student background and school environment variables. When properly taken account of in the analysis, the complex probability sample design used to collect NAEP data provides a rigorous justification for extending survey results to the entire U.S. student population. Recent developments in the analysis of categorical data from complex surveys provide a straight forward methodology for taking account of sample design through the proper estimation of subpopulation proportions (domain P-values) and their covariance matrices [Koch, Freeman, and Freeman (1975)]. These vectors of subpopulation P-values are then fit to linear models in the domain defining variables using the sample design based covariance matrix to calculate weighted least squares fits. Wald statistics that take the form of Hotelling's multivariate T^2 statistic are then used to test for the goodness of fit of the model and to subsequently test for the significance of model effects. The following chapter surveying theoretical results begins with a section on the Wald statistic/weighted least squares theory for testing hypotheses about NAEP domain p-values.

While the weighted least squares methodology provides a straight forward solution to the NAEP p-value analysis problem, the computer software required to properly estimate sample design based covariance matrices is not generally available to NAEP data users. Section 2.2 summarizes methods explored by Fellegi (1980) and Rao and Scott (1981) for adjusting test statistics derived from standard statistical software packages. The standard statistical software packages employ either implicitly or explicitly covariance matrix calculations appropriate for simple random samples. The adjustments proposed by Fellegi, Rao and Scott involve dividing the standard chi-squared statistic by a generalized design effect (Deff) summarizing the ratio of sample design based P-value variances and covariances divided by their respective simple random sampling variances and covariances. To display the sample design effects of multi-stage clustering, stratification, and unequal weighting on the generalized Deff, a design effect model identity is developed in Section 2.3 for the P-value covariance matrix and for linear contrasts among sample P-values. The proper sample design based inference for NAEP balanced fits obtained as dummy variable regression coefficients is presented in Section 2.4 along with the analogous generalized Deff adjustment. Since much of the descriptive analysis of NAEP data utilizes subpopulation averages across several individual item P-values, the proper covariance matrix estimation and generalized Deff adjustment methods for such statistics are explored in Section 2.5.

2. THEORETICAL RESULTS

2.1 The Wald Statistic/Weighted Least-Square Theory

To illustrate the weighted least-squares approach to the analysis of NAEP domain P-values, consider a single questionnaire item with response categories labelled r = 1,2,...,R. Suppose further that there are D subpopulation domains of interest labelled d = 1,2,...,D. These domains can be viewed as the student subpopulations formed by a cross-classification of background variables including Race, Sex, Region, Type of Community, and Parents Education. Let $X_d(t)$ depict a subpopulation indicator variable taking the value 1 when student t belongs to domain d and zero otherwise. Similarly, let $Y_r(t)$ denote a response indicator variable taking the value 1 when student t gives response r to the specified item and zero otherwise. With the U.S. student population size for a particular age class denoted by M, the population count of domain d members giving response r is

$$Y(dr) = \sum_{t=1}^{M} X_d(t)\, Y_r(t) .$$

The U.S. student population size for subpopualtion d is specified as

$$X(d) = \sum_{t=1}^{M} X_d(t) .$$

The proportion of domain d members giving response r is then defined as the ratio

$$P(dr) = Y(dr)/X(d) .$$

The NAEP sample estimates for these student subpopulation response proportions (domain P-values) have the form

$$\hat{P}(dr) = [\sum_{k=1}^{m} W(k)\ X_d(k)\ Y_r(k)] \div [\sum_{k=1}^{m} W(k)\ X_d(k)]$$

where m denotes the number of sample students and W(k) is a sample weight incorporating the reciprocal of the sample student k inclusion probability $\pi(k)$ and various adjustments for sample school and student nonresponse.

To simplify the notation, consider the column vector $\hat{\underset{\sim}{P}}(d)$ of (R-1) subpopulation d response proportions consisting of the first (R-1) estimates defined above. Stacking these domain specific vectors on top of one another, a single column vector $\hat{\underset{\sim}{P}}$ with D(R-1) elements is produced. Let $\hat{\underset{\sim}{V}}_p(DES)$ denote the D(R-1) by D(R-1) estimated covariance matrix derived for $\hat{\underset{\sim}{P}}$ according to one of the three asympotically equivalent methods of variance estimation for nonlinear statistics from complex probability samples, namely the Taylor Series linearization (TSL) method, balanced repeated replication (BRR), and jackknife replication (JKR). Krewski and Rao (1978) have established limiting conditions for the asymptotic equivalence of these covariance matrix estimation methods. A central limit theorem establishing conditions for the asymptotic normality of studentized P-values

$$t(dr) = [\hat{P}(dr) - P(dr)] / [\hat{V}_{DES}(dr)]^{\frac{1}{2}}$$

is also presented by Krewski and Rao when any of the three linearization methods is used to approximate the sample design based variance $V_{DES}(dr)$ for the ratio statistic $\hat{P}(dr)$. Such a central limit theorem provides the theoretical justification for assuming that the vector $\hat{\underset{\sim}{P}}$ of estimated domain P-values will be distributed approximately as a D(R-1)- variate normal vector with mean $\underset{\sim}{P}$ and covariance matrix $\underset{\sim}{V}_p(DES)$.

Assuming that the conditions for asymptotic normality apply, weighted least squares methods following Grizzle, Starmer, and Koch (1969) can be used to fit linear models to functions of the estimated domain P-values. Let $\underset{\sim}{G}(\underset{\sim}{P})$ denote a vector of A continuous, linearly independent functions with partial derivatives through second order. Define the A by D(R-1) matrix of partial derivatives with row a denoted by

$$\underset{\sim}{H}_a(\underset{\sim}{P}) \; = \; \partial G_a / \partial \underset{\sim}{P} \; = \; (\partial G_a/\partial P_1, \; \ldots, \; \partial G_a/\partial P_{D(R-1)}).$$

The matrix of these partial derivatives is evaluated at the estimated $\hat{\underset{\sim}{P}}$ values to define

$$\hat{\underset{\sim}{H}} \; = \; \underset{\sim}{H} \, (\hat{\underset{\sim}{P}}) \; .$$

Consider, for example, the logit function where

$$G_a(\underset{\sim}{P}) \; = \; \log_e \, [P_a/(1-P_a)]$$

with a = 1,2,...D indexing the D domain P-values associated with the typical R=2 (correct-incorrect) item response breakdown. For this logistic function

$$\hat{\underset{\sim}{H}}_a \; = \; (0,0,\ldots,1/\hat{P}_a(1-\hat{P}_a), \; . \; . \; . \; 0) \; .$$

Now, with $\hat{\underset{\sim}{G}} = \underset{\sim}{G}(\hat{\underset{\sim}{P}})$ denoting the sample estimate for the vector of A functions, one notes that $\hat{\underset{\sim}{G}}$ is asymptotically A-variate normal with mean vector $\underset{\sim}{G}$ = G(P) and asymptotic covariance matrix

$$\underset{\sim}{V}_G(\mathrm{DES}) = [H \; \underset{\sim}{V}_P(\mathrm{DES}) \; H^T].$$

where H^T denotes the transpose of H. A consistent estimate for $\underset{\sim}{V}_G$(DES) is

$$\hat{\underset{\sim}{V}}_G(\mathrm{DES}) \; = \; [\hat{H} \; \hat{\underset{\sim}{V}}_P(\mathrm{DES}) \; \hat{H}^T]$$

One can now proceed to fit a general linear model of the form

$$\underset{\sim}{G}(\underset{\sim}{P}) \; = \; X \, \underset{\sim}{\beta}$$

where the columns of X specify selected main effect and interaction contrasts in terms of the domain defining Race, Sex, Type of Community and

Parents Education variables. The asymptotically efficient BAN (Best Asymptotically Normal) estimator for the coefficient vector $\underset{\sim}{\beta}$ is then

$$\hat{\underset{\sim}{\beta}} = [X^T \hat{\underset{\sim}{V}}_G(DES)^{-1} X]^{-1} X^T \hat{\underset{\sim}{V}}_G(DES)^{-1} \hat{\underset{\sim}{G}}$$

with asymptotic covariance matrix

$$\hat{V}_\beta(DES) = [X^T \hat{\underset{\sim}{V}}_G(DES)^{-1} X]^{-1} .$$

To test the fit of the model, that is H_0: $\underset{\sim}{G}(\underset{\sim}{P}) = X \underset{\sim}{\beta}$, the residual quadratic form

$$T^2(Fit) = (\hat{\underset{\sim}{G}} - \underset{\sim}{X} \hat{\underset{\sim}{\beta}})^T \hat{\underset{\sim}{V}}_G(DES)^{-1} (\hat{\underset{\sim}{G}} - \underset{\sim}{X} \hat{\underset{\sim}{\beta}})$$

is a Wald (1943) statistic which has the form of Hotelling's multivariate T^2 statistic. Asymptotically $T^2(Fit)$ is distributed as Chi-Square (χ^2) with degrees of freedom df = rank of $\underset{\sim}{X}$ under the null hypothesis. For subasymptotic situations where the number of replicates used to form the covariance matrix estimator does not substantially exceed the rank of X, a transformation of T^2 to Snedecor's F may be appropriate. This leads to

$$F(df, L - df + 1) = (L - df + 1)\ T^2(Fit)/df(L)$$

where L is the number of degrees of freedom suggested by the quadratic form used to estimate $\hat{\underset{\sim}{V}}_P(DES)$ and df is the rank of $\underset{\sim}{X}$. This transformed Wald Statistic is compared against critical values of Snedecor's F with df numerator and L - df + 1 denominator degrees of freedom.

Failing to reject H_0: $G(\underset{\sim}{P}) = X \underset{\sim}{\beta}$, one can entertain linear hypotheses of the form H_0: $\underset{\sim}{C}\, \underset{\sim}{\beta} = \underset{\sim}{\Phi}$, with $\underset{\sim}{\Phi}$ denoting a null vector. The associated Wald Statistic is

$$T^2(\underset{\sim}{C}) = (\underset{\sim}{C}\, \hat{\underset{\sim}{\beta}})^T [\underset{\sim}{C}\, \hat{\underset{\sim}{V}}_\beta(DES)\, \underset{\sim}{C}^T]^{-1} (\underset{\sim}{C}\, \hat{\underset{\sim}{\beta}})$$

which is asymptotically χ^2 (rank of C) under the null hypothesis. The F transformed alternative

$$F(c, L - c + 1) = (L - c + 1)\ T^2(\underset{\sim}{C})/cL$$

is compared with critical values of Senedecor's F with c = rank of C numerator degrees of freedom and (L - c + 1) denominator degrees of freedom. In the next section the approximate methods of Rao and Scott (1979) based on simple random sampling covariance matrices $\hat{V}_P(SRS)$ and generalized design effect (deff) adjustments are explored.

2.2 Generalized Design Effect Methods

Rao and Scott (1981) considered the asymptotic distribution of Wald Statistics based on SRS covariance matrices for testing general hypotheses of the form

$$H_0:\ G_a(\underset{\sim}{P}) = 0, \qquad a = 1,2,\ldots A$$

where $\underset{\sim}{P}$ can be viewed as the national response distribution for a specified NAEP exercise. This vector of universe level proportions corresponds to the $\underset{\sim}{P}(d)$ domain specific item response proportions introduced earlier with P(dr) denoting the domain or subpopulation proportion selecting the coded response option r = 1,2,...,(R-1). Since the P(dr) sum to one over all R mutually exclusive response levels, only (R-1) of the parameters is required to fully characterize the response distribution. While Rao and Scott's results focus on the single population problem, they can be extended simply to the multiple subpopulation or domain problem by allowing $\underset{\sim}{P}$ to represent the extended vector

$$\underset{\sim}{P}^T = < \underset{\sim}{P}^T(1),\ .\ .\ .,\ \underset{\sim}{P}^T(d),\ .\ .\ .,\ \underset{\sim}{P}^T(D) > .$$

Letting

$$\underset{\sim}{H}_a(\underset{\sim}{P}) = \partial G_a(\underset{\sim}{P})/\partial \underset{\sim}{P}$$

$$= [\partial G_a(\underset{\sim}{P})/\partial P(11),\ .\ .\ .\ .\ .\ .\ .,\ \partial G_a(\underset{\sim}{P})/\partial P(D,R-1)]$$

denote the vector of partial derivatives of $G_a(\underset{\sim}{P})$ with respect to the D(R-1) elements of $\underset{\sim}{P}$, the population of response proportions, then

$$\chi^2_{SRS}(\hat{G}) = \hat{\underset{\sim}{G}}^T \; [\hat{\underset{\sim}{H}} \; \hat{V}_P(SRS) \; \hat{\underset{\sim}{H}}^T]^{-1} \; \hat{\underset{\sim}{G}}$$

is the SRS based Wald statistic for testing the hypothesis that the vector of A functions $\underset{\sim}{G}^T = [G_1(\underset{\sim}{P}), \ldots, G_A(\underset{\sim}{P})]$ are simultaneously zero with $\hat{\underset{\sim}{H}}$ denoting the matrix of all partial derivatives $\underset{\sim}{H}(\underset{\sim}{P})$ evaluated at $\underset{\sim}{P} = \hat{\underset{\sim}{P}}$. Interest in the SRS based χ^2 statistic stems from the simple computational form for $\hat{V}_p(SRS)$. The simple random sampling covariance matrix for $\hat{\underset{\sim}{P}}$ is approximated as a block diagonal matrix with (R-1) by (R-1) blocks of the form

$$\hat{V}_d(SRS) = \{\text{diag } [\hat{\underset{\sim}{P}}(d)] - \hat{\underset{\sim}{P}}(d) \; \hat{\underset{\sim}{P}}^T(d)\}/m(d)$$

where m(d) is the domain d sample size and diag $[\hat{\underset{\sim}{P}}(d)]$ is the (R-1) by (R-1) diagonal matrix with diagonal elements $\hat{P}(dr)$. This SRS based covariance matrix is formed simply from the weighted domain P-values and the observed domain sample sizes. For the typical NAEP analysis of correct responses with R = 2,

$$\hat{V}_d(SRS) = \hat{P}(d) \; [1-\hat{P}(d)]/m(d)$$

and $\hat{V}_p(SRS)$ is a D by D diagonal matrix with the $\hat{V}_d(SRS)$ quantities on the diagonal. Under the null hypothesis H_0: $G(\underset{\sim}{P}) = \Phi$,

$$\chi^2_{SRS}(G) \approx \sum_{a=1}^{A} \delta_{oa} \; \chi^2_a$$

where the δ_a's are the eigenvalues of

$$[H \; V_p(SRS) \; H^T]^{-1} \; [H \; V_p(DES) \; H^T] \; ,$$

$\delta_1 \geq \ldots \geq \delta_A > 0$, the χ^2_a's are independent χ^2_1 (single degree of freedom chi-squared) random variables and δ_{oa} is the value of δ_a under H_0.

Rao and Scott point out that the δ_a's can be interpreted as design effects of linear combinations L_a of the components of $H \; \hat{\underset{\sim}{P}}$. Letting λ_a

denote the a-th eigenvalue of

$$V_p(SRS)^{-1}\ V_p(DES)$$

then the δ eigenvalues can be bounded by the λ eigenvalues as follows:

$$\lambda_a \geq \delta_a \geq \lambda_{D(R-1)}$$

for $a = 1, \ldots, A$, since the L_a are particular linear combinations of the $\hat{P}(dr)$'s. Using a result in Anderson and Das Gupta (1963), Rao and Scott establish more precise bounds for the δ_a in terms of the λ_a; namely,

$$\lambda_a \geq \delta_a \geq \lambda_{D(R-1)-A+a}\ .$$

This inequality is useful for specifying an alternative to the following χ^2 test statistic proposed by Rao and Scott (R&S),

$$\chi^2_{R\&S}\ (\underset{\sim}{G}) \ = \ \chi^2_{SRS}\ (\underset{\sim}{G})\ /\ \hat{\delta}_.$$

$$\approx \sum_{a=1}^{A}\ [\delta_{oa}/\delta_{o.}]\ \chi^2_a$$

where $\hat{\delta}_.$ denotes an estimate of the mean eigenvalue $\delta_.$ with the $\hat{\delta}_a$ depicting eigenvalues of

$$[\hat{H}\ \hat{V}_p(SRS)\ H^T]^{-1}\quad [\hat{H}\ \hat{V}_p(DES)\ \hat{H}^T]\ .$$

Since the estimation of the $\hat{\delta}_a$ and associated $\hat{\delta}_.$ require knowledge of the full design based covariance matrix, there is no real utility in using this approximation when one could just as well use the appropriate design based Wald statistic. Using the sharp bounds for δ_a, one notes that $\delta_.$ lies between the average of the A largest λ_a's and the mean of the A smallest λ_a's. This implies that $\delta_.$ should get close to $\lambda_.$ as the number of G_a functions (A) approaches D(R-1). With $A = D(R-1)$ independent G_a functions such that H is nonsingular it is clear that $\delta_. = \lambda_.$. If the λ_a's show little variation, so that the mean of the A largest and A smallest λ's are similar, then one can also expect that $\hat{\delta}_. \doteq \hat{\lambda}$. The advantage of using $\hat{\lambda}_.$

instead of $\hat{\delta}_.$ to adjust Chi-Square $X^2{}_{SRS}(G)$ is the ease of estimating $\hat{\lambda}_.$. We note that

$$\hat{V}_P(SRS) \;=\; \text{BLK-DIAG}\;[\{\text{diag}\;[\underset{\sim}{\hat{P}}(d)] - \underset{\sim}{\hat{P}}(d)\;\underset{\sim}{\hat{P}}(d)'\}/m(d)]$$

is block diagonal with blocks comprised of the domain specific multinomial covariance matrices $\hat{V}_{P(d)}(SRS)$. Therefore, extending results of Rao and Scott one obtains

$$\begin{aligned}
\hat{\lambda}_. \;&=\; \text{trace}\;\{\hat{V}_P(SRS)^{-1}\;\;\hat{V}_P(DES)\}\;/\;D(R-1)\\
&=\; \sum_{d=1}^{D}\;\text{trace}\;\hat{V}_{P(d)}(SRS)^{-1}\;\hat{V}_{P(d)}(DES)\;/\;D(R-1)\\
&=\; \sum_{d=1}^{D}\;\sum_{r=1}^{R-1}\;\hat{V}(dr|DES)\;m(d)\;/\;\hat{P}(dr)\;D(R-1)\\
&=\; \sum_{d=1}^{D}\;\sum_{r=1}^{R-1}\;[1-\hat{P}(dr)]\;\;\widehat{DEFF}\;(dr)\;/\;D(R-1)
\end{aligned}$$

where

$$\widehat{DEFF}(dr) \;=\; \hat{V}\;(dr|DES)\;/\;\{\hat{P}(dr)\;[1-\hat{P}(dr)]\;/\;m(d)\}$$

is the design effect for the cell proportion $\hat{P}(dr)$.

Returning to the NAEP correct-incorrect response pattern (R=2), $\hat{\lambda}_.$ simplifies to

$$\hat{\lambda}_. \;=\; \sum_{d=1}^{D}\;\widehat{DEFF}\;(dr)\;/\;D,$$

the mean of the domain specific design effects. This result follows from the diagonal form for

$$\hat{V}_P(SRS)^{-1} \;=\; \text{diag}\;[m(d)/\{\hat{P}(d)\;[1-\hat{P}(d)]\}]\;.$$

In either case, the generalized design effect $\hat{\lambda}_.$ is a simple function of domain P-value design effects. When the design effects for subpopulation P-values, $\hat{P}(dr)$, are published then $\hat{\lambda}_.$ can be formed without knowledge of the design based covariances between response porportions $\hat{P}(dr)$ and $\hat{P}(d'r')$ from different subpopulations ($d \neq d'$). In the following chapter, numerical comparisons of the adjusted SRS based χ^2 statistics

$$\chi^2_{ADJ}\,(\underset{\sim}{G}) \;=\; \chi^2_{SRS}\,(\underset{\sim}{G}) \,/\, \hat{\lambda}\ .$$

and the associated design based $\chi^2_{DES}(\underset{\sim}{G})$ Wald statistics are explored.

To model the effects of sample design features like stratification clustering and unequal weighting on the $\chi^2_{SRS}(G)$ statistic, one can develop a model for the generalized design effect matrix

$$V_P(SRS)^{-1} \quad V_P(DES)\ .$$

Consider, for example, a two-stage design with S primary frame units (PFU's), say schools, with M(s) secondary units (students) in the s-th PFU such that $\sum_{s=1}^{S} M(s) = M$. A with replacement selection of n primary sampling units (PSU's) is first made with single draw probabilities $\phi(s) = [M(s)/M]$. A subsequent with replacement simple random sample of m second stage units is then drawn from each sample PSU. For a single universal domain, Rao and Scott display the following partitioning for

$$V_P(DES) = V_P(SRS) + (m-1) \sum_{s=1}^{S} \phi(s)\ (\underset{\sim}{P}_s - \underset{\sim}{P})(\underset{\sim}{P}_s - \underset{\sim}{P})^T/nm$$

$$= V_P(SRS)\ [I + (m-1)R]$$

where $\phi(s) = M(s)/M$ is the fraction of all students who attend school s;

$$\underset{\sim}{P}^T_s = [P_s(1),\ .\ .\ .,\ P_s(R-1)]$$

is the vector of (R-1) response option proportions for students in school s, and

$$R = \{\text{diag}\ (\underset{\sim}{P}) - \underset{\sim}{P}\underset{\sim}{P}^T\}^{-1} \sum_{s=1}^{S} \phi(s)(\underset{\sim}{P}_s - \underset{\sim}{P})(\underset{\sim}{P}_s - \underset{\sim}{P})'$$

is the matrix analogue of the intra-cluster correlation coefficient with $\underset{\sim}{P}$ denoting the universe level vector of (R-1) response option proportions. With this partitioning one can show that the eigenvalues of

$$V_P(SRS)^{-1}V_P(DES) = [I + (m-1)R]$$

have the form

$$\lambda_a = [1 + (m-1)\,\rho_a]$$

where ρ_a is the a-th largest eigenvalue of the intracluster correlation matrix R. Rao and Scott call these ρ_a quantities generalized measures of homogeneity, analogous to the intracluster correlation ρ. For the simple goodness-of-fit hypothesis H_o: $\underset{\sim}{P} = \underset{\sim}{P}_0$, the simple random sample (SRS) χ^2 can be written as

$$\chi^2_{SRS}\,(P_0) = \sum_{a=1}^{(R-1)} [1+(m-1)\rho_{oa}]\,\chi^2_a$$

where the χ^2_a are single degree of freedom central Chi-square variables. For a portable value of $\hat{\rho}_.$ useful for modeling $\hat{\lambda}_. = [1+(m-1)\hat{\rho}_.]$ in comparable samples with differing cluster sizes m, one could use

$$\hat{\rho}_. = (\hat{\lambda}_.-1)/(m-1)$$

$$= [\{\sum_{r=1}^{R-1} [1-\hat{P}(r)]\widehat{DEFF}(r)/(R-1)\}-1]/(m-1).$$

An extension of this self-weighting, two stage, with replacement model for λ_a to a NAEP type design with effects for unequal weighting, stratification, and clustering is presented in the following section. These results are used to display the effect of sample design features on SRS based Wald Statistics for testing the fit of linear models

$$H_o\text{:}\quad G(\underset{\sim}{P}) = X\beta$$

and linear hypotheses regarding the model parameters β; that is,

$$H_o: \ C\beta = \Phi.$$

2.3 Design Effect Models for P-Value Covariance Matrices

To develop a design effect model for the covariance matrix of a vector $\underset{\sim}{P}$ of D National Assessment domain P-values, we consider a three stage design with n county sized primary sampling units selected from a universe of N such units where the random frequency of selection for primary frame unit PFU(ℓ) is n(ℓ). These n(ℓ) are akin to the λ(ℓ) random selection indicators for without replacement samples where λ(ℓ) is 1 when PFU(ℓ) belongs to the sample and zero otherwise. The n(ℓ) are allowed to assume values greater than 1 so as to accomodate so-called self-representing or certainty units. Following Chromy (1979), one can use these random selection frequencies to characterize a class of probability proportional to size (PPS) selection schemes including PPS with replacement, PPS without replacement, and PPS minimum replacement (PMR). The PPS nature of these selection schemes implies that

$$\begin{aligned} E\{n(\ell)\} = En(\ell) &= ns(\ell)/s(+) \\ &= n\phi(\ell) \end{aligned}$$

where s(ℓ) is a size measure known for each primary frame unit PFU(ℓ) and

$$s(+) = \sum_{\ell=1}^{N} s(\ell)$$

is the universe level aggregate size measure. For the NAEP design, the size measure s(ℓ) is typically the estimated PFU enrollment for the 13-year-old target population. For with replacement selections, the n(ℓ) are multinomial frequencies with

$$E\{n(\ell)n(\ell')\} = n(n-1)\phi(\ell)\phi(\ell')$$

when $\ell \neq \ell'$. For without replacement PPS designs $n(\ell) = \lambda(\ell)$, the zero-one selection indicator, and

$$E\{n(\ell)n(\ell')\} = \pi(\ell\ell')$$

with $\pi(\ell\ell')$ denoting the joint inclusion probability for the frame units PFU(ℓ) and PFU(ℓ'). For Chromy's probability minimum replacement (PMR) design

$$|n(\ell)-En(\ell)|=|n(\ell)-n\phi(\ell)| < 1.$$

Specifically, for PMR designs

$$Pr\{n(\ell) = Int[n\phi(\ell)]+1\} = Frac\ [n\phi(\ell)]$$

and

$$Pr\{n(\ell) = Int[n\phi(\ell)]\} = 1-Frac\ [n\phi(\ell)]$$

where Int(x) denotes the integer part of x and Frac(x) depicts the fractional part of x. The PMR feature allows multiple selection from certainty units with $n\phi(\ell)>1$ such that the number of hits $n(\ell)$ is derived by randomly rounding the $En(\ell) = n\phi(\ell)$ proportional to size allocations up or down. The sampling variance function for this class of selection schemes has been parameterized in terms of variance and covariance components by Folsom (1980) utilizing double draw probabilities

$$\phi(\ell\ell') = \begin{cases} E\{n(\ell)[n(\ell)-1]\}/n(n-1) & \text{if } \ell=\ell' \\ E\{n(\ell)n(\ell')\}/n(n-1) & \text{if } \ell\neq\ell' . \end{cases}$$

These double draw probabilities $\phi(\ell\ell')$ and the associated single draw probabilities $\phi(\ell)$ were dervied by Folsom as the expectations of single draw sampling unit indicators $\lambda_{\ell}(i)$ that take the value 1 when $n(\ell)>0$ and selected primary frame unit PFU(ℓ) is randomly assigned primary sampling unit (PSU) label i with i ranging from 1 to n; otherwise $\lambda_\ell(i) = 0$. With the labels assigned as a random permutation of the digits 1, . . ., n, one can show that

$$E\{\lambda_\ell(i)\} = \phi(\ell) \text{ for all } i$$

and

$$E\{\lambda_\ell(i)\lambda_{\ell'}(i')\} = \phi(\ell\ell') \text{ for all } i \neq i' \ ,$$

where expectation is taken over repeated samples and repeated random PSU label assignments. The single draw indicators have the additional properties

$$\sum_{\ell=1}^{N} \lambda_\ell(i) = 1$$

for all i, and

$$\sum_{\ell=1}^{n} \lambda_\ell(i) = n(\ell).$$

These results lead to the following $\phi(\ell)$ and $\phi(\ell\ell')$ summation identities:

$$\sum_{\ell=1}^{N} \phi(\ell) = 1$$

and

$$\begin{aligned} \sum_{\ell=1}^{N} \phi(\ell\ell') &= E\{\lambda_{\ell'}(i) \sum_{\ell=1}^{N} \lambda_\ell(i)\} \\ &= E\{\lambda_{\ell'}(i)\} \\ &= \phi(\ell'). \end{aligned}$$

In the following subsection, the single draw indicators are used to define unbiased single draw variates in terms of the following single PFU ratio type estimators

$$\underset{\sim}{y}(\ell) = s(+)\ \underset{\sim}{Y}(\ell)/s(\ell)$$

$$= \underset{\sim}{Y}(\ell)/\phi(\ell)$$

where

$$\underset{\sim}{Y}(\ell)^T = [Y_{11}(\ell), \ . \ . \ ., \ Y_{dr}(\ell), \ . \ . \ ., \ Y_{D(R-1)}(\ell)]$$

denotes a row vector of D(R-1) PFU totals and the superscript T denotes matrix transposition. In the illustration above, $Y_{dr}(\ell)$ will denote the (dr)-th element of the vector $\underset{\sim}{Y}(\ell)$ specifying the number of age eligible domain d students attending school in PFU(ℓ) who would give response option r to a particular NAEP item.

2.3.1 Single Stage Covariance Matrix Models

To develop a design effect representation for the covariance matrix of a three-stage design statistic we begin by developing single stage results. Extending Folsom's (1980) single stage results to vector valued statistics, we consider the corresponding vector valued single draw variate

$$\underset{\sim}{y}(i) = \sum_{\ell=1}^{N} \lambda_\ell(i)\ \underset{\sim}{y}(\ell),$$

with $\underset{\sim}{y}(\ell)$ denoting the single PFU ratio type estimator $\underset{\sim}{Y}(\ell)/\phi(\ell)$ defined previously.

Now, one can show that each single draw sample variate y(i) is an unbiased estimator of the universe total $\underset{\sim}{Y}(+)$; that is,

$$E\{\underset{\sim}{y}(i)\} = \sum_{\ell=1}^{N} E\{\lambda_\ell(i)\}\ \underset{\sim}{y}(\ell)$$

$$= \sum_{\ell=1}^{N} \underset{\sim}{Y}(\ell)$$

$$= \underset{\sim}{Y}(+) \text{ for all } i.$$

Similarly, the covariance matrix for each $\underset{\sim}{y}(i)$ is

$$E\{[\underset{\sim}{y}(i)-\underset{\sim}{Y}(+)][\underset{\sim}{y}(i)-\underset{\sim}{Y}(+)]^T\} = E\{\sum_{\ell=1}^{N} \lambda_\ell(i)[\underset{\sim}{y}(\ell)-\underset{\sim}{Y}(+)]\}\ \{\sum_{\ell'=1}^{N} \lambda_{\ell'}(i)[\underset{\sim}{y}(\ell')-\underset{\sim}{Y}(+)]^T\}$$

$$= E\{\sum_{\ell=1}^{N} \lambda_\ell(i)[\underset{\sim}{y}(\ell)-\underset{\sim}{Y}(+)][\underset{\sim}{y}(\ell)-\underset{\sim}{Y}(+)]^T\}$$

since

$$\lambda_\ell(i)\ \lambda_{\ell'}(i) = \begin{matrix} \lambda_\ell(i) & \text{when } \ell = \ell' \\ 0 & \text{when } \ell \neq \ell'. \end{matrix}$$

Taking the expectation over repeated samples and PSU label assignments (E) inside the summation, one obtains

$$\Sigma_y(\mathrm{PSU}) = \sum_{\ell=1}^{N} \phi(\ell)[\underset{\sim}{y}(\ell)-\underset{\sim}{Y}(+)][\underset{\sim}{y}(\ell)-\underset{\sim}{Y}(+)]^T.$$

The cross-covariance matrix between $\underset{\sim}{y}(i)$ and $\underset{\sim}{y}(i')$ is derived similarly as

$$\Sigma R_y(\mathrm{PSU}) = E\{\sum_{\ell=1}^{N} \sum_{\ell'=1}^{N} \lambda_\ell(i)\lambda_{\ell'}(i')[\underset{\sim}{y}(\ell)-\underset{\sim}{Y}(+)][\underset{\sim}{y}(\ell')-\underset{\sim}{Y}(+)]^T\}$$

$$= \sum_{\ell=1}^{N} \sum_{\ell'=1}^{N} \phi(\ell\ell')[\underset{\sim}{y}(\ell)-\underset{\sim}{Y}(+)][\underset{\sim}{y}(\ell')-\underset{\sim}{Y}(+)]^T$$

for all $i \neq i'$. The fact that the single drawn sample variates have common covariance and cross-covariance matrices, provides a simple classical derivation of the variance for the mean of the single draw variate vectors

$$\bar{\underset{\sim}{y}} = \sum_{i=1}^{n} \underset{\sim}{y}(i)/n$$

$$= \sum_{\ell=1}^{N} \left[\sum_{i=1}^{n} \lambda_{\ell}(i) \right] \underset{\sim}{Y}(\ell)/n\phi(\ell)$$

$$= \sum_{\ell=1}^{N} n(\ell)\underset{\sim}{Y}(\ell)/En(\ell).$$

Notice that in the recast version $\bar{y}$ becomes the standard Hansen-Hurwitz (1943) estimator for a PPS with replacement selection. For a without replacement PPS sample with $En(\ell)=\pi(\ell)$, $\bar{\underset{\sim}{y}}$ is the Horvitz-Thompson (1952) estimator. For the intermediate PMR designs, $\bar{\underset{\sim}{y}}$ is Chromy's (1979) estimator. In terms of the common covariance and cross-covariance matrices $\Sigma_y(PSU)$ and $\Sigma R_y(PSU)$, the covariance matrix for $\bar{\underset{\sim}{y}}$ is

$$V_{\bar{y}}(PSU) = \Sigma_y(PSU)/n + (n-1)\Sigma R_y(PSU)/n$$

$$= \Sigma_y(PSU)\ [I+(n-1)R_y(PSU)]/n$$

with

$$R_y(PSU) = \Sigma_y(PSU)^{-1}\ \Sigma R_y(PSU)$$

defining the cross-correlation matrix and I denoting the D(R-1) by D(R-1) identity matrix.

The following alternative expressions for $\Sigma_y(PSU)$ and $\Sigma R_y(PSU)$ make it easy to see that the form for $V_{\bar{y}}(PSU)$ developed above is equivalent to the Yates-Grundy (1953) type variance expression presented in Chromy for this general class of designs:

$$\Sigma_y(PSU) = \sum_{\ell=1}^{N} \sum_{\ell'\neq\ell}^{N} \phi(\ell)\phi(\ell')[\underset{\sim}{y}(\ell)-\underset{\sim}{y}(\ell')][\underset{\sim}{y}(\ell)-\underset{\sim}{y}(\ell')]^{T}/2$$

and

$$\Sigma R_y(PSU) = \sum_{\ell=1}^{N} \sum_{\ell'\neq\ell}^{N} [\phi(\ell)\phi(\ell')-\phi(\ell\ell')][\underset{\sim}{y}(\ell)-\underset{\sim}{y}(\ell')][\underset{\sim}{y}(\ell)-\underset{\sim}{y}(\ell')]^{T}/2\ .$$

Weighting these component matrices together as indicated in $V_{\bar{y}}(\mathrm{PSU})$ leads to

$$V_{\bar{y}}(\mathrm{PSU}) = \sum_{\ell=1}^{N} \sum_{\ell' \neq \ell}^{N} [\phi(\ell)\phi(\ell') - (n-1)\phi(\ell\ell')/n][\underset{\sim}{y}(\ell) - \underset{\sim}{y}(\ell')][\underset{\sim}{y}(\ell) - \underset{\sim}{y}(\ell')]^T/2$$

$$V_{\bar{y}}(\mathrm{PSU}) = \sum_{\ell=1}^{N} \sum_{\ell' \neq \ell}^{N} [En(\ell)En(\ell') - En(\ell)n(\ell')] \left\{ \frac{\underset{\sim}{Y}(\ell)}{En(\ell)} - \frac{\underset{\sim}{Y}(\ell')}{En(\ell')} \right\} \left\{ \frac{\underset{\sim}{Y}(\ell)}{En(\ell)} - \frac{\underset{\sim}{Y}(\ell')}{En(\ell')} \right\}^T/2.$$

When the PSU selection scheme gives all pairs of primary frame units ℓ and ℓ' a chance of being represented in the sample so that $En(\ell)n(\ell') > 0$ for all $\ell \neq \ell'$, the alternative component matrix expressions suggest unbiased estimators

$$\hat{\Sigma}_y(\mathrm{PSU}) = \sum_{\ell=1}^{N} \sum_{\ell' \neq \ell} [n(\ell)n(\ell')/En(\ell)n(\ell')]\phi(\ell)\phi(\ell')\underset{\sim}{\delta}(\ell\ell')\underset{\sim}{\delta}(\ell\ell')^T/2$$

and

$$\hat{\Sigma}R_y(\mathrm{PSU}) = \sum_{\ell=1}^{N} \sum_{\ell' \neq \ell} [n(\ell)n(\ell')/En(\ell)n(\ell')][\phi(\ell)\phi(\ell') - \phi(\ell\ell')]\underset{\sim}{\delta}(\ell\ell')\underset{\sim}{\delta}(\ell\ell')^T/2$$

with

$$\underset{\sim}{\delta}(\ell\ell') = [\underset{\sim}{y}(\ell) - \underset{\sim}{y}(\ell')].$$

For a multi-stage sample such as the NAEP design, the PFU vector totals $\underset{\sim}{y}(\ell)$ imbeded in the definition of our $\underset{\sim}{\delta}(\ell\ell')$ quantities must be estimated based on second and subsequent stages of sampling. The unbiased estimation of stage specific component matrices is complicated by this process. The following section develops the three stage covariance matrix model for the vector valued total estimator $\underset{\sim}{\hat{Y}}(+)$.

2.3.2 Three-Stage Covariance Matrix Models for Estimated Totals

For a three-stage NAEP type design where c schools are selected for a given package in each sample PSU and m students are selected for package assignment in each sample school, random selection frequencies $\xi(\ell s)$ and $m(\ell st)$ characterize the number of selections of school s in PFU(ℓ) and the number of selections of student t in school (ℓs). For NAEP sample designs, the school and student level selections are without replacement. The school selections are made with probability proportional to estimated age class enrollment, say $A(\ell s)$; that is

$$\begin{aligned} E\{\xi(\ell s)|n(\ell)=1\} &= c\ A(\ell s)/A(\ell +) \\ &= c\ \phi(s|\ell). \end{aligned}$$

When multiple hits are allowed on the primary frame units, $n(\ell) > 1$ independent replicated samples of c schools are drawn from PFU(ℓ).

The second stage conditional double draw probabilities are defined as

$$\begin{aligned} E\{\lambda_{\ell s}(ij)\lambda_{\ell s'}(ij')|\lambda_{\ell}(i)=1\} &= \begin{array}{l} E\{\xi(\ell s)[\xi(\ell s)-1]\}/c(c-1) \text{ if } s=s' \\ E\{\xi(\ell s)\xi(\ell s')\}/c(c-1) \text{ if } s\neq s' \end{array} \\ &= \phi\ (ss'|\ell) \end{aligned}$$

where $\lambda_{\ell s}(ij)$ is 1 if $\xi(\ell s)>0$ and school s in PFU(ℓ) is randomly assigned sample school label (ij) given $\lambda_{\ell}(i)=1$. The third stage sample is a simple random selection without replacement so that

$$\begin{aligned} E\{m(\ell st)|\xi(\ell s) > 0\} &= m/M(\ell s) \\ &= m\phi(t|\ell s) \end{aligned}$$

with equal single draw probabilities $\phi(t|\ell s) = M(\ell s)^{-1}$ where $M(\ell s)$ denotes the number of age eligible students in school list unit SCH(ℓs). The conditional double draw probabilities at the third stage are

$$\begin{aligned} E\{\lambda_{\ell st}(ijk)\lambda_{\ell st'}(ijk')|\lambda_{\ell}(i)\lambda_{\ell s}(ij) =1\} &= 1/M(\ell s)[M(\ell s)-1] \text{ if } t\neq t' \\ &= \phi(tt'|\ell s) \end{aligned}$$

and otherwise $\phi(tt|\ell s) = 0$ since for without replacement selections $\lambda_{\ell st}(ijk)\ \lambda_{\ell st}(ijk') = 0$; that is, the student list unit t can not be labeled both sample student k and k' since student list unit t can be selected only once. With these definitions, a three stage single draw variate is defined as

$$\underset{\sim}{y}(ijk) = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} \alpha_{\ell st}(ijk)\ \underset{\sim}{Y}(\ell st)/\phi(\ell st)$$

with the three stage single draw indicator $\alpha_{\ell st}(ijk)$ defined as the product of the stagewise indicators

$$\alpha_{\ell st}(ijk) = \lambda_{\ell}(i)\ \lambda_{\ell s}(ij)\ \lambda_{\ell st}(ijk)\ .$$

The corresponding three stage single draw probability is defined analogously; that is,

$$E\{\alpha_{\ell st}(ijk)\} = \phi(\ell st) = \phi(\ell)\ \phi(s|\ell)\ \phi(t|\ell s).$$

With these definitions, it is not difficult to see that

$$\begin{aligned} E\ \{\underset{\sim}{y}(ijk)\} &= \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} \underset{\sim}{Y}\ (\ell st) \\ &= \underset{\sim}{Y}(+++) \end{aligned}$$

for all sample students STU(ijk) where $\underset{\sim}{Y}(+++)$ is the universe total of the response vector $\underset{\sim}{Y}(\ell st)$ with D(R-1) elements of the form

$$Y_{dr}(\ell st) = X_d(\ell st)\ Y_r(\ell st)$$

where $X_d(\ell st)$ takes the value one when student list unit SLU(ℓst) belongs to subpopulation domain d and zero otherwise. The covariance matrices for these three stage single draw variables can be derived simply using conditional expectations. For example, consider

$$E\{[\underset{\sim}{y}(ijk) - \underset{\sim}{Y}(+++)]\ [\underset{\sim}{y}(ijk) - \underset{\sim}{Y}(+++)]^T\} = \Sigma_y\ (PSU) + \Sigma_y\ (SCH) + \Sigma_y(STU)$$

$$= \mathrm{Cov}\ [\underset{\sim}{y}(ijk)]\ ,$$

where

$$\Sigma_y\ (STU) = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \phi(\ell s) \sum_{t=1}^{M(\ell s)} \phi(t|\ell s)\ [\underset{\sim}{y}(\ell st) - \underset{\sim}{y}(\ell s_{.})]\ [\underset{\sim}{y}(\ell st) - \underset{\sim}{y}(\ell s_{.})]^T$$

with the school list unit mean $\underset{\sim}{y}(\ell s.)$ defined as

$$\underset{\sim}{y}(\ell s_{.}) = \sum_{t=1}^{M(\ell s)} \phi(t|\ell s)\ \underset{\sim}{y}(\ell st)$$

$$= \underset{\sim}{Y}\ (\ell s+)\ /\ \phi(\ell)\ \phi(s|\ell)$$

$$= \underset{\sim}{Y}\ (\ell s+)\ /\ \phi(\ell s)\ .$$

For NAEP type designs with simple random selections at the third stage one obtains the simplified form

$$\Sigma_y\ (STU) = M(++)^2 \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \left[\ \frac{\theta^2(\ell s)}{\phi(\ell s)}\ \right] \Sigma_y(\ell s)$$

where $\theta(\ell s) = M(\ell s)/M(++)$ denotes the school list unit (ℓs) fraction of the total student population count $M(++)$. The covariance matrix for school list unit (ℓs) is

$$\Sigma_y\ (\ell s) = [\mathrm{diag}\ \{\underset{\sim}{\pi}(\ell s)\} - \underset{\sim}{\pi}(\ell s)\ \underset{\sim}{\pi}(\ell s)^T]$$

with

$$\underset{\sim}{\pi}(\ell s)^T = \sum_{t=1}^{M(\ell s)} \underset{\sim}{Y}(\ell st)^T / M(\ell s)$$

$$= [\pi_{11}(\ell s),\ \ldots,\ \pi_{dr}(\ell s),\ \ldots,\ \pi_{D(R-1)}(\ell s)]$$

denoting the vector of $D(R-1)$ cell proportions for school list unit (ℓs). The dr-th cell in our example denotes membership in domain d and item response group r.

The school stage covariance matrix component has the form

$$\Sigma_y(\mathrm{SCH}) = \sum_{\ell=1}^{N} \phi(\ell) \sum_{s=1}^{S(\ell)} \phi(s|\ell)\ [\underset{\sim}{y}(\ell s.)-\underset{\sim}{y}(\ell..)]\ [\underset{\sim}{y}(\ell s_.) - \underset{\sim}{y}(\ell..)]^T$$

with

$$[\underset{\sim}{y}(\ell s_.) - \underset{\sim}{y}(\ell_{..})] = \{\underset{\sim}{Y}(\ell s+)/\phi(\ell)\ \phi(s|\ell) - \underset{\sim}{Y}(\ell++)/\phi(\ell)\}$$
$$= M(++)\ \{[\frac{\theta(\ell s)}{\phi(\ell s)}]\ \underset{\sim}{\pi}(\ell s) - [\frac{\theta(\ell)}{\phi(\ell)}]\ \underset{\sim}{\pi}(\ell)\}$$

where $\theta(\ell) = M(\ell+)/M(++)$ denotes the fraction of student age eligibles in PFU(ℓ). If the relative size measure

$$\phi(\ell s) = [S(\ell)/S(+)]\ [A(\ell s)/A(\ell+)]$$
$$= M(\ell s)/M(++),$$

then the sample is self-weighting since

$$\phi(\ell st) = \phi(\ell s)/M(\ell s) = 1/M(++)$$

and the total inclusion probability for student list unit SLU (ℓst) is

$$\pi(\ell st) = ncm\ \phi(\ell st) = ncm/M(++),$$

a constant for all nsm sample students. In this simplified case

$$\Sigma_y\ (\mathrm{STU}) = M(++)^2 \sum_{\ell=1}^{N} \sum_{s=1}^{M(\ell s)} \theta(\ell s)\ \{\mathrm{diag}\ [\underset{\sim}{\pi}(\ell s)] - \underset{\sim}{\pi}(\ell s)\ \underset{\sim}{\pi}(\ell s)^T\}$$

and

$$\Sigma_y\ (\mathrm{SCH}) = M(++)^2 \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \theta(\ell s)\ [\underset{\sim}{\pi}(\ell s) - \underset{\sim}{\pi}(\ell)]\ [\underset{\sim}{\pi}(\ell s) - \underset{\sim}{\pi}(\ell)]^T.$$

The general form of the PSU level covariance component matrix is

$$\Sigma_y\ (\mathrm{PSU}) = \sum_{\ell=1}^{N} \phi(\ell)\ [\underset{\sim}{y}(\ell_{..}) - \underset{\sim}{y}(\ldots)]\ [\ \underset{\sim}{y}(\ell_{..}) - \underset{\sim}{y}(\ldots)]^T$$

with

$$[\underset{\sim}{y}(\ell_{..}) - \underset{\sim}{y}(\ldots)] = M(++)\ \{\ [\frac{\theta(\ell)}{\phi(\ell)}\ ]\ \underset{\sim}{\pi}(\ell) - \underset{\sim}{\pi}\ \}\ .$$

For a self-weighting sample

$$\Sigma_y \text{ (PSU)} = M^2(++) \sum_{\ell=1}^{N} \theta(\ell)\ [\underset{\sim}{\pi}(\ell)-\underset{\sim}{\pi}]\ [\underset{\sim}{\pi}(\ell) - \underset{\sim}{\pi}]^T.$$

In general, one notes that

$$\begin{aligned}
\text{Cov}[\underset{\sim}{y}(ijk)] &= \Sigma_y \text{ (PSU)} + \Sigma_y \text{ (SCH)} + \Sigma_y \text{ (STU)} \\
&= \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} \phi(\ell st)\ [\underset{\sim}{y}(\ell st) - \underset{\sim}{y}(...)]\ [\underset{\sim}{y}(\ell st) - \underset{\sim}{y}(...)]^T \\
&= \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} \underset{\sim}{Y}(\ell st)\ \underset{\sim}{Y}(\ell st)^T/\phi(\ell st) - \underset{\sim}{Y}(+++)\ Y(+++)^T \\
&= M^2(++)\ \{ \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} \underset{\sim}{Y}(\ell st)\ \underset{\sim}{Y}(\ell st)^T/M(++)^2\ \phi(\ell st) - \underset{\sim}{\pi}\ \underset{\sim}{\pi}^T\} \\
&= \Sigma_y^2(\text{TOT}).
\end{aligned}$$

When the sample is self-weighting with

$$\phi(\ell st) = 1/M(++),$$

the common covariance matrix for each $\underset{\sim}{y}(ijk)$ is

$$\Sigma_y \text{ (TOT)} = M^2(++)\ \{\text{diag } [\underset{\sim}{\pi}] - \underset{\sim}{\pi}\ \underset{\sim}{\pi}^T\}.$$

Notice that in the self-weighting case, $\Sigma_y(\text{TOT})/M(++)^2$ is the SRS with replacement multinomial covariance component matrix.

Various cross-covariance components can be define for the three-stage single draw variables. These cross-covariance components are derived as follows:

$$\begin{aligned}
\text{Cov}\ [\underset{\sim}{y}(ijk);\ \underset{\sim}{y}(ijk')] &= \underset{\text{PSU}}{\text{Cov}}\ [y(i..)] + \underset{\text{PSU}}{E}\ \underset{\text{SCH}}{\text{Cov}}\ [\underset{\sim}{y}(ij.)] \\
&\quad + \underset{\text{PSU}}{E}\ \underset{\text{SCH}}{E}\ \underset{\text{STU}}{\text{Cov}}\ [\underset{\sim}{y}(ijk);\ \underset{\sim}{y}(ijk')] \\
&= \Sigma_y \text{ (PSU)} + \Sigma_y \text{ (SCH)} + \Sigma R_y \text{ (STU)}
\end{aligned}$$

where

$$\underset{\sim}{y}(ij.) = \underset{STU}{E}\ \{\underset{\sim}{y}(ijk)|PSU,\ SCH\}$$

denotes the conditional expectation of y(ijk) over repeated student selections and label assignments given the PSU and school selections and label assignments. The conditional expectation over school and student selection and label assignment of $\underset{\sim}{y}(ijk)$ given the PSU selection and labelling is similarly denoted by

$$\underset{\sim}{y}(i..) = \underset{SCH}{E}\ \underset{STU}{E}\ \{\underset{\sim}{y}(ijk|PSU\}.$$

The matrices Σ_y (PSU) and Σ_y (SCH) were defined previously, and

$$\Sigma R_y\ (STU) = \sum_{\ell=1}^{N}\ \sum_{s=1}^{S(\ell)}\ \phi(\ell s)\ \sum_{t=1}^{M(\ell s)}\ \sum_{t'=1}^{M(\ell s)}\ \underset{\sim}{\delta}(\ell st)\underset{\sim}{\delta}(\ell st)^T/M(\ell s)[M(\ell s)-1]$$

with

$$\underset{\sim}{\delta}(\ell st) = [\underset{\sim}{y}(\ell st) - \underset{\sim}{y}(\ell s_{.})]$$

$$= M(\ell s)\ [\underset{\sim}{Y}(\ell st) - \underset{\sim}{\pi}(\ell s)]/\phi(\ell s).$$

With further manipulation the following form for the between student within school cross-covariance matrix is obtained

$$\Sigma R_y\ (STU) = -M(++)^{2}\ \sum_{\ell=1}^{N}\ \sum_{s=1}^{S(\ell)}\ [\ \frac{\theta^2(\ell s)}{\phi(\ell s)}\ ]\ \Sigma_y\ (\ell s)/[M(\ell s)-1]$$

recalling that

$$\Sigma_y(\ell s) = \{\mathrm{diag}\ [\underset{\sim}{\pi}(\ell s)] - \underset{\sim}{\pi}(\ell s)\ \underset{\sim}{\pi}(\ell s)^T\}\ .$$

Turning to the between school within PSU cross-covariance matrix one obtains

$$Cov\ [\underset{\sim}{y}(ijk);\ \underset{\sim}{y}(ij'k')] = \underset{PSU}{Cov}\ [\underset{\sim}{y}(i_{..})] + \underset{PSU}{E}\ \underset{SCH}{Cov}\ [\underset{\sim}{y}(ij_{.});\ \underset{\sim}{y}(ij'_{.})]$$

$$= \Sigma_y\ (PSU) + \Sigma R_y\ (SCH)$$

with

$$\Sigma R_y(SCH) = \sum_{\ell=1}^{N}\ \phi(\ell)\ \sum_{s=1}^{S(\ell)}\ \sum_{s'=1}^{S(\ell)}\ \phi(ss'|\ell)\ \underset{\sim}{\delta}(\ell s)\ \underset{\sim}{\delta}(\ell s')^T$$

where

$$\underset{\sim}{\delta}(\ell s) = M(++)\ \{[\frac{\theta(\ell s)}{\phi(\ell s)}]\ \underset{\sim}{\pi}(\ell s) - [\frac{\theta(\ell)}{\phi(\ell)}]\ \underset{\sim}{\pi}(\ell)\}.$$

Finally, the between PSU cross-covariance matrix is defined as follows

$$\begin{aligned} \text{Cov}\ [\underset{\sim}{y}(ijk);\ \underset{\sim}{y}(i'j'k')] &= \sum_{\ell=1}^{N} \phi(\ell\ell')\ \underset{\sim}{\delta}(\ell)\ \underset{\sim}{\delta}(\ell')^T \\ &= \Sigma R_y(\text{PSU}) \end{aligned}$$

where

$$\begin{aligned} \underset{\sim}{\delta}(\ell) &= [\underset{\sim}{y}(\ell_{..}) - \underset{\sim}{y}(_{...})] \\ &= \{\ [\frac{\theta(\ell)}{\phi(\ell)}]\ \underset{\sim}{\pi}(\ell) - \underset{\sim}{\pi}\}. \end{aligned}$$

Armed with the covariance and cross-covariance component definitions specified above, one can derive the covariance matrix for

$$\bar{\underset{\sim}{y}} = \sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} \underset{\sim}{y}(ijk)/ncm,$$

the three-stage analogue of the general class of with replacement and PMR single stage estimators. The covariance matrix for $\bar{\underset{\sim}{y}}$ is

$$\begin{aligned} \text{Cov}\ [\bar{\underset{\sim}{y}}] = {} & \Sigma_y(\text{PSU})\ [I + (n-1)\ R_y(\text{PSU})]/n + \Sigma_y(\text{SCH})\ [I + (c-1)\ R_y(\text{SCH})]/nc \\ & + \Sigma_y\ (\text{STU})\ [I + (m-1)\ R_y\ (\text{STU})]/ncm. \end{aligned}$$

A design effect version of this model can be formed as follows:

$$\begin{aligned} \text{Cov}[\bar{\underset{\sim}{y}}] = {} & \{\Sigma_y(\text{SRS})/ncm\}\ \bar{W}\ \{I + (ncm-1)\ \Delta_y(\text{PSU})\ R_y(\text{PSU}) \\ & + (cm-1)\ [\Delta_y(\text{PSU})\ \{I - R_y(\text{PSU})\} + \Delta_y(\text{SCH})\ R_y(\text{SCH})] \\ & + (m-1)\ [\Delta_y(\text{SCH})\ \{I - R_y(\text{SCH})\} + \Delta_y(\text{STU})\ R_y(\text{STU})]\} \end{aligned}$$

where

$$\begin{aligned} \bar{W} &= \Sigma_y(\text{SRS})^{-1}\ \Sigma_y(\text{TOT}) \\ &\equiv \text{The unequal weighting effect;} \end{aligned}$$

$\Delta_y(PSU) = \Sigma_y(TOT)^{-1}\ \Sigma_y(PSU)$

$\equiv$ The within PSU clustering effect;

$R_y(PSU) = \Sigma_y(PSU)^{-1}\ \Sigma R_y(PSU)$

$\equiv$ The PSU stratification and PMR selection effect;

$\Delta_y(SCH) = \Sigma_y(TOT)^{-1}\ \Sigma_y(SCH)$

$\equiv$ The within school clustering effect;

$R_y(SCH) = \Sigma_y(SCH)\ \Sigma R_y(SCH)$

$\equiv$ The school stratification and PMR selection effect;

$\Delta_y(STU) = \Sigma_y(TOT)^{-1}\ \Sigma_y(STU)$

$= [I - \Delta_y(PSU) - \Delta_y(SCH)]$

$\equiv$ Between student within school fraction of total variation;

and

$R_y(STU) = \Sigma_y(STU)^{-1}\ \Sigma R_y(STU)$

$\equiv$ The effect of without replacement student selection.

As suggested above, the cross-covariance matrices account for both the effects of minimum replacement (PMR) or without replacement selections and stratification. The effect of explicit and implicit stratification is expressed through the sampling expectation of selection frequency products, $E\ n(\ell)\ n(\ell')$, and associated double draw probabilities, $\phi(\ell\ell')$. Recalling the Yates-Grundy form of the PSU level variance function

$$V_{\bar{y}}(PSU) = \sum_{\ell=1} \sum_{\ell' \neq \ell} [En(\ell)En(\ell') - En(\ell)n(\ell')]\ \underset{\sim}{d}(\ell\ell')\ \underset{\sim}{d}(\ell\ell')^T/2$$

with

$$\underset{\sim}{d}(\ell\ell') = \{\underset{\sim}{Y}(\ell)/En(\ell) - \underset{\sim}{Y}(\ell')/En(\ell')\}\ ,$$

it is clear that explicit stratification would imply the independence of selection frequencies $n(\ell)$ and $n(\ell')$ for PFU's in different strata. This independence would cause the between PSU contrasts $\underset{\sim}{d}(\ell\ell')$ in $V_{\bar{y}}(PSU)$ to drop out since $En(\ell)n(\ell') = En(\ell)\ En(\ell')$. Therefore, while the $V_{\bar{y}}(PSU)$ variance expression and the variance-covariance component analogue are not written in the familiar stratified form, they reduce to such a form when the $[En(\ell)\ En(\ell') - En(\ell)n(\ell')]$ coefficients are set to zero for the between stratum terms. With Chromy's sequential PMR zone selection scheme, implicit stratification effects are achieved by purposively ordering the frame listing so that proximate units are expected to be more alike than distant units in terms of the survey outcome measures. The reduction in sampling variance associated with the effect of implicit stratification is reflected in the Yates-Grundy variance form by a tendency for the coefficients

$$[En(\ell)\ En(\ell') - En(\ell)n(\ell')]$$

to approach zero as the distance between frame units ℓ and ℓ' increases. The variance-covariance component representation for $V_{\bar{y}}(PSU)$ displays the combined effect of minimum replacement selection and implicit stratification in the form $[1 + (n-1)\ R_y(PSU)]$. For a scalar statistic this expression reduces to $[1 + (n-1)\rho_y(PSU)]$ where $\rho_y(PSU)$, the common correlation among the n single draw variates $y(i)$, is expected to be increasingly negative as the efficacy of the implicit stratification improves.

2.3.3 Design Effect Model for the NAEP P-Value Covariance Matrix

The design effect model developed for the linear statistic $\bar{\underset{\sim}{y}}$, the vector of estimated domain by item response category totals $\tilde{Y}(dr)$, can be extended to the vector $\underset{\sim}{P}$ of $D(R-1)$ response category proportions considered previously by applying the Taylor Series linearization technique implicit in the section 2.2 treatment of generalized design effect methods. Begin by letting

$$G_{dr}[\underset{\sim}{Y}(+++)] = Y(dr)/\sum_{r=1}^{R} Y(dr) = P(dr)$$

and let $\underset{\sim}{H}_{dr}$ denote the 1 × D(R-1) vector with elements

$$h_{dr}(uv) = \partial P(dr)/\partial Y(uv) \ .$$

The elements of $\underset{\sim}{H}_{dr}$ have the following form

$$\underset{\sim}{H}_{dr} = \begin{array}{ll} [1 - P(dr)]/X(d) & \text{if } u = d \text{ and } v = r \\ - P(dr)/X(d) & \text{if } u = d \text{ and } v \neq r \\ 0 \quad \text{otherwise} & \end{array}$$

where

$$X(d) = \sum_{r=1}^{R} Y(dr) \ .$$

Letting $\underset{\sim}{H}^T = (\underset{\sim}{H}_{11}, \ldots, \underset{\sim}{H}^T_{dr}, \ldots, \underset{\sim}{H}^T_{D(R-1)})$ defining the matrix of partial derivatives of $\underset{\sim}{P}$ with respect to the elements of $\underset{\sim}{Y}(+++)$ where the T superscript denotes matrix transposition, then

$$V_p(DES) = \underset{\sim}{H} \ \mathrm{Cov}[\bar{\underset{\sim}{y}}] \ \underset{\sim}{H}^T \ .$$

The expression for the design based covariance matrix of $\underset{\sim}{P}$ stated above can now be used along with the three-stage component representation for Cov $[\bar{\underset{\sim}{y}}]$ to produce an analogous component representation for V_p(DES). This is accomplished by defining analogous covariance and cross-variance matrices for each stage as follows:

$$\Sigma_p(STAGE) = \underset{\sim}{H} \ \Sigma_y(STAGE) \ \underset{\sim}{H}^T$$

and

$$\Sigma R_p(STAGE) = \underset{\sim}{H} \ \Sigma R_y(STAGE) \ \underset{\sim}{H}^T$$

with STAGE assuming the PSU, SCH, and STU levels for the NAEP design.

An alternative form of the Taylor series linearization that provides explicit definitions for the component matrices is to define linearized variates

$$\underset{\sim}{z}(ijk) = \underset{\sim}{H}\;\underset{\sim}{y}(ijk)$$

$$= \sum_{\ell=1}^{N}\;\sum_{s=1}^{S(\ell)}\;\sum_{t=1}^{M(\ell s)} \alpha_{\ell st}(ijk)\; z(\ell st)$$

where

$$\underset{\sim}{z}(\ell st) = \underset{\sim}{H}\;\underset{\sim}{y}(\ell st)$$

$$= \underset{\sim}{H}\;\underset{\sim}{Y}(\ell st)/\phi(\ell st)$$

$$= M(\ell s)\;\underset{\sim}{Z}(\ell st)/\phi(\ell)\phi(s|\ell)\;.$$

The $\underset{\sim}{Z}(\ell st) = \underset{\approx}{HY}(\ell st)$ vectors defined above have elements of the form

$$Z_{dr}(\ell st) = X_d(\ell st)\;[Y_r(\ell st) - P(dr)]/X(d)$$

recalling that $X_d(\ell st)$ is the one-zero indicator for domain d membership and $Y_r(\ell st)$ is the one-zero indicator for response category r. Using the linearized three-stage single draw variates $\underset{\sim}{z}(\ell st)$ in place of the $\underset{\sim}{y}(\ell st)$ vectors in the $\Sigma_y(\;)$ and $\Sigma R_y(\;)$ definitions yields $\Sigma_z(\;)$ and $\Sigma R_z(\;)$ matrices such that

$$\Sigma_p(\mathrm{STAGE}) = \Sigma_z(\mathrm{STAGE}) = \underset{\sim}{H}\;\Sigma_y(\mathrm{STAGE})\;\underset{\sim}{H}^T$$

and

$$\Sigma R_p(\mathrm{STAGE}) = \Sigma R_z(\mathrm{STAGE}) = \underset{\sim}{H}\;\Sigma_y(\mathrm{STAGE})\;\underset{\sim}{H}^T\;.$$

In terms of these quantities, the school level population mean vectors

$$\underset{\sim}{z}(\ell s_{\bullet}) = \sum_{t=1}^{M(\ell s)} \underset{\sim}{z}(\ell st)/M(\ell s)$$

have elements

$$z_{dr}(\ell s_{\bullet}) = X_d(\ell s+)[P_{dr}(\ell s) - P(dr)]/X(d)\;\phi(\ell s)$$

$$= \theta_d(\ell s)\;[P_{dr}(\ell s) - P(dr)]/\theta(\ell s)$$

where $P_{dr}(\ell s)$ is the proportion of domain d members of school list unit (ℓs) that would give item response option r and $\theta_d(\ell s)$ is the fraction of all domain d members attending school in school list unit (ℓs). The PSU level mean vector

$$\underset{\sim}{z}(\ell..) = \sum_{s=1}^{S(\ell)} \phi(s|\ell)\underset{\sim}{z}(\ell s.)$$

has elements

$$z_{dr}(\ell..) = X_d(\ell{+}{+})\,[P_{dr}(\ell) - P(dr)]/X(d)\phi(\ell)$$
$$= \theta_d(\ell)\,[P_{dr}(\ell) - P(dr)]/\phi(\ell)\ .$$

If one lets $D_\zeta(\ell s)$ denote a D(R-1) by D(R-1) diagonal matrix with the (dr)-th element $\zeta_{dr}(\ell s) = \theta_d(\ell s)/\phi(\ell s)$, then the school level mean vector of linearized variates is

$$\underset{\sim}{z}(\ell s.) = D_\zeta(\ell s)\,[\underset{\sim}{P}(\ell s) - \underset{\sim}{P}]\ .$$

Defining $D_\zeta(\ell)$ with (dr)-th element $\zeta_{dr}(\ell) = \theta_d(\ell)/\phi(\ell)$, a similar form for the PSU level mean vector is obtained, namely

$$\underset{\sim}{z}(\ell..) = D_\zeta(\ell)[\underset{\sim}{P}(\ell)-\underset{\sim}{P}]\ .$$

When no members of domain d attend school list unit SCH(ℓs), then $\theta_d(\ell s) = 0$ and $P_{dr}(\ell s) = 0$. Similarly, if no members of domain d attend school in primary frame unit PFU(ℓ), then $\theta_d(\ell) = 0$ and $P_{dr}(\ell) = 0$.

In terms of these linearized variate vectors, the stage specific covariance matrices have the form

$$\Sigma_P(STU) = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \phi(\ell s) \sum_{t=1}^{M(\ell s)} [\underset{\sim}{z}(\ell st)-\underset{\sim}{z}(\ell s.)]\,[\underset{\sim}{z}(\ell st)-\underset{\sim}{z}(\ell s.)]^T/M(\ell s)$$

$$= \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \phi(\ell s) \sum_{t=1}^{M(\ell s)} \underset{\sim}{z}(\ell st)\underset{\sim}{z}(\ell st)^T/M(\ell s)$$

$$- \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \phi(\ell s)\underset{\sim}{z}(\ell s.)\underset{\sim}{z}(\ell s.)^T\ .$$

Letting $D_\theta(\ell st)$ denote a D(R-1) by D(R-1) diagonal matrix with elements $\theta_d(\ell st) = X_d(\ell st)/X(d)$, the between student covariance matrix becomes

$$\Sigma_P(\mathrm{STU}) = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} D_\theta(\ell st)[\underset{\sim}{Y}(\ell st)-\underset{\sim}{P}]\,[\underset{\sim}{Y}(\ell st)-\underset{\sim}{P}]^T D_\theta(\ell st)\ M(\ell s)/\phi(\ell s)$$

$$- \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \phi(\ell s) D_\zeta(\ell s)[\underset{\sim}{P}(\ell s)-\underset{\sim}{P}][\underset{\sim}{P}(\ell s)-\underset{\sim}{P}]^T D_\zeta(\ell s)$$

where the (dr)-th element of $\underset{\sim}{Y}(\ell st)$ is $Y_r(\ell st)$, the one-zero indicator for item response category r. The between school within PSU covariance matrix is

$$\Sigma_P(\mathrm{SCH}) = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \phi(\ell s)[\underset{\sim}{z}(\ell s.)-\underset{\sim}{z}(\ell..)][\underset{\sim}{z}(\ell s.)-\underset{\sim}{z}(\ell..)]^T$$

$$= \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \phi(\ell s)\underset{\sim}{z}(\ell s.)\underset{\sim}{z}(\ell s.)^T$$

$$- \sum_{\ell=1}^{N} \phi(\ell) z(\ell..) z(\ell..)^T .$$

Therefore

$$\Sigma_P(\mathrm{SCH}) = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \phi(\ell s) D_\zeta(\ell s)[\underset{\sim}{P}(\ell s)-\underset{\sim}{P}]\,[\underset{\sim}{P}(\ell s)-\underset{\sim}{P}]^T D_\zeta(\ell s)$$

$$- \sum_{\ell=1}^{N} \phi(\ell) D_\zeta(\ell)[\underset{\sim}{P}(\ell)-\underset{\sim}{P}]\,[\underset{\sim}{P}(\ell)-\underset{\sim}{P}]^T D_\zeta(\ell) .$$

The between PSU covariance matrix is

$$\Sigma_P(\mathrm{PSU}) = \sum_{\ell=1}^{N} \phi(\ell)\ [\underset{\sim}{z}(\ell..)-\underset{\sim}{z}(...)]\ [\underset{\sim}{z}(\ell..)-\underset{\sim}{z}(...)]^T$$

$$= \sum_{\ell=1}^{N} \phi(\ell)\ \underset{\sim}{z}(\ell..)\underset{\sim}{z}(\ell..)^T$$

since

$$\underset{\sim}{z}(...) = \sum_{\ell=1}^{N} \phi(\ell)\ \underset{\sim}{z}(\ell..)$$

$$= \sum_{\ell=1}^{N} \phi_d(\ell)\ [\underset{\sim}{P}(\ell) - \underset{\sim}{P}] = \underset{\sim}{0}\ .$$

Therefore, with the ϕ weighted mean of the linearized vectors $\underset{\sim}{z}$ equivalent to the null vector, the between PSU covariance matrix can be written as

$$\Sigma_P(\text{PSU}) = \sum_{\ell=1}^{N} \phi(\ell) D_\zeta(\ell)\ [\underset{\sim}{P}(\ell)-\underset{\sim}{P}]\ [\underset{\sim}{P}(\ell)-\underset{\sim}{P}]^T D_\zeta(\ell)\ .$$

Combining these results, one obtains the following expression for the total covariance matrix

$$\text{Cov}\ [\underset{\sim}{z}(ijk)] = \Sigma_P(\text{TOT})$$

$$= \Sigma_P(\text{PSU}) + \Sigma_P(\text{SCH}) + \Sigma_P(\text{STU})$$

$$= \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} D_\theta(\ell st)[\underset{\sim}{Y}(\ell st)-\underset{\sim}{P}]\ [\underset{\sim}{Y}(\ell st)-\underset{\sim}{P}]^T D_\theta(\ell st) M(\ell s)/\phi(\ell s)\ .$$

Letting $W(\ell st) = M(\ell s)/[ncm\phi(\ell s)]$ denote the sample weight or inverse selection probability for student (ℓst), one can recast $\Sigma_P(\text{TOT})$ as a block diagonal matrix with blocks of the form

$$\Sigma_P^d(\text{TOT}) = ncm\ \{ \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} W(\ell st) X_d(\ell st)[\underset{\sim}{Y}(\ell st)-\underset{\sim}{P}(d)]\ [\underset{\sim}{Y}(\ell st)-\underset{\sim}{P}(d)]^T/X(d)^2\}$$

$$= ncm\ \bar{W}(d)\bar{f}(d)\{\text{diag}\ [\underset{\sim}{P}W(d)]-\underset{\sim}{P}W(d\ \underset{\sim}{P}(d)^T-\underset{\sim}{P}(d)\underset{\sim}{P}W(d)^T+\underset{\sim}{P}(d)\underset{\sim}{P}(d)^T\}/Em(d)$$

where

$$\bar{W}(d) = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} W(\ell st) X_d(\ell st)/X(d)$$

is the average weight for all domain d members in the universe and

$$\bar{f}(d) = Em(d)/X(d)$$

is the expected sampling fraction for domain d with

$$\mathrm{Em}(d) = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} [\mathrm{ncm}\ \phi(\ell s)/M(\ell s)] X_d(\ell st)$$

$$= \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} W(\ell st)^{-1}\ X_d(\ell st)$$

denoting the expected domain d sample size over repeated samples. The $\underset{\sim}{P}W(d)$ vectors represent the weighted universe level response option distribution for domain d members; that is

$$\underset{\sim}{P}W(d) = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} W(\ell st)\ X_d(\ell st)\underset{\sim}{Y}(\ell st)/W_d(+++)$$

with $W_d(+++)$ denoting the universe level weight sum for domain d members. For a self-weighting sample with common weight

$$W(\ell st) = M(\ell s)/\mathrm{ncm}\ \phi(\ell s)$$

$$= M(++)/\mathrm{ncm}$$

one observes that

$$\bar{W}(d) = M(++)/\mathrm{ncm}$$

$$\bar{f}(d) = \mathrm{ncm}/M(++)$$

and

$$\underset{\sim}{P}W(d) = \underset{\sim}{P}(d)\ .$$

For a self-weighting design one notes therefore that $\Sigma_P^d(\mathrm{TOT})/\mathrm{nsm}$ defined on page 34 assumes the with replacement simple random sampling multinomial form; that is

$$\Sigma_P^d(\mathrm{TOT})/\mathrm{ncm} = \{\mathrm{diag}\ [\underset{\sim}{P}(d)] - \underset{\sim}{P}(d)\underset{\sim}{P}(d)^T\}/\mathrm{Em}(d)$$

with the expected domain sample size specified as

$$\mathrm{Em}(d) = \mathrm{ncm}\ X(d)/M(++)$$

$$= \mathrm{ncm}\ \pi(d)\ .$$

For the typical multi-stage PPS sample design utilizing approximate size measures, the unequal weighting effect is defined as

$$\bar{\omega}_P = V_P(SRS)^{-1} \Sigma_P(TOT)/ncm \ .$$

The matrix $\bar{\omega}_P$ is block diagonal with blocks

$$\bar{\omega}\ (d) = \bar{W}(d)\ \bar{f}(d)\{\text{diag}\ [P(d)]-P(d)P(d)^T\}^{-1}\ \{\ \text{diag}\ [PW(d)]-PW(d)P(d)$$
$$-\ \underset{\sim}{P}(d)\underset{\sim}{P}W(d)^T + \underset{\sim}{P}(d)\underset{\sim}{P}(d)^T\ \}$$

When a correct-incorrect dichotomous response distribution is considered, the domain d effect of unequal weighting $\bar{\omega}_P(d)$ can be recast in the following form

$$\bar{\omega}_P(d) = \bar{W}(d)\bar{f}(d)\ [PW(d)-2PW(d)P(d)+P(d)^2]/P(d)[1-P(d)]$$
$$= \bar{W}(d)\bar{f}(d)\ \{[PW(d)/P(d)] + [1-PW(d)]/[1-P(d)] - 1\}\ .$$

Recalling that $\bar{f}(d)$ is the subpopulation d mean of the inclusion probabilities $\pi(\ell st) = W(\ell st)^{-1}$, the product of the average weight $\bar{W}(d)$ and the expected domain d sampling fraction $\bar{f}(d)$ can be written as

$$\bar{W}(d)\ \bar{f}(d) = \bar{W}(d) \div \bar{W}_H(d)$$

where

$$\bar{W}_H(d) = 1/\bar{f}(d)$$
$$=[\sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} W(\ell st)^{-1}\ X_d(\ell st)/X(d)]^{-1}$$

is the universe level harmonic mean of the $W(\ell st)$ weights for domain d members. Noting that the $W(\ell st)$ weights are nonnegative quantities, it is clear that $\bar{W}(d)\bar{f}(d) \geq 1$ since the harmonic mean $\bar{W}(d)$ is less than the arithmetic mean $\bar{W}(d)$ for nonnegative variables.

2.3.4 Estimates for Composite Covariance Matrix Components

To produce estimates for the three-stage covariance matrix components defined in the previous section, one can begin by building a consistent estimator for $\Sigma_p(TOT)$. Note that

$$\hat{\bar{W}}(d) = \frac{\left[\sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} X_d(ijk)\ W(ijk)/\pi(ijk) \right]}{\left[\sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} X_d(ijk)\ W(ijk) \right]}$$

is a consistent ratio estimator for $\bar{W}(d)$ and that

$$\hat{\bar{f}}(d) = m(d) \div \left[\sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} X_d((ijk)\ W(ijk) \right]$$

is similarly a consistent estimator for $\bar{f}(d)$. Therefore

$$\hat{\bar{W}}(d)\hat{\bar{f}}(d) = m(d)\ \{ \frac{\left[\sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} X_d(ijk)\ W(ijk)^2 \right]}{\left[\sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} X_d(ijk)\ W(ijk) \right]^2}$$

is a consistent estimator for the associated unequal weighting factor. The consistent sample estimator $\hat{\bar{W}}(d)\hat{\bar{f}}(d)$ is equivalent to the unequal weighting design effect proposed by Kish (1965) and others. The derivation presented here shows that there is an additional term in the unequal weighting effect that contrasts the universe level weighted mean of the correct response indicator, PW(d), with the subpopulation proportion correct P(d). This additional unequal weighting factor

$$Q(d) = \{[PW(d)/P(d)] + [1-PW(d)]/[1-P(d)] - 1\}$$

is less than one when $P(d) \geq 0.5$ and $PW(d) > P(d)$. When $PW(d) > P(d)$, this implies that $\bar{W}_1(d) > \bar{W}_0(d)$ where $\bar{W}_1(d)$ denotes the universe level mean of the weights for domain d members who respond correctly and $\bar{W}_0(d)$ is the corresponding mean for domain d members who respond incorrectly. Therefore $Q(d) < 1$ when the harmonic mean of the inclusion probabilities for domain d members who respond incorrectly is greater than the harmonic mean inclusion probability for those domain d members who would respond correctly; that is for $P(d) \geq 0.5$, $Q(d) < 1$ when the sample design overrepresents domain d members who would respond incorrectly to the item. For items with $P(d) < 0.5$, $Q(d)$ is less than one when $PW(d) < P(d)$ which implies overrepresentation of domain d members who would respond correctly. For the NAEP design where schools in low income inner city areas are overrepresented, there will be a tendency for overrepresentation of persons who would respond incorrectly. The effect of this overrepresentation on the $Q(d)$ quantities should not be expected to counterbalance the rather substantial unequal weighting design effects. The total population value of $\hat{\bar{W}}\hat{\bar{f}}$ is around 1.35 for a single NAEP package sample. Consider for example an item with $P = 0.55$ and $PW = 0.95$, then $Q = 0.838$ and $\bar{\omega}_p = 1.13$. Consistent sample estimates of the $PW(d)$ can be formed using the squared weights $W(ijk)^2$ to compute the weighted proportion giving response option r as

$$\hat{P}W(dr) = \left\{ \frac{\sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} W(ijk)^2 \; X_d(ijk) \; Y_r(ijk)}{\sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} W(ijk)^2 \; X_d(ijk)} \right\} .$$

Utilizing these consistent estimators for $\bar{W}(d)$, $\bar{f}(d)$, and $PW(dr)$, a consistent estimator for the domain d block of $\Sigma_p(TOT)/ncm$ is

$$\hat{W}_p^d(TOT)/ncm = \hat{\bar{W}}(d)\hat{\bar{f}}(d) \; S\hat{W}_p(d)/m(d)$$

where

$$S\hat{W}_p(d) = \{ \text{diag} \, [P\hat{\underset{\sim}{W}}(d)] - P\hat{\underset{\sim}{W}}(d)\hat{\underset{\sim}{P}}(d)^T \;\; \hat{\underset{\sim}{P}}(d)P\hat{\underset{\sim}{W}}(d)^T + \hat{\underset{\sim}{P}}(d)\hat{\underset{\sim}{P}}(d)^T \} .$$

These considerations lead to the consistent estimator for $\Sigma_p(TOT)$

$$\hat{\Sigma}_p(TOT) = ncm \; \text{BLK-DIAG} \; \{\hat{\bar{W}}(d)\hat{\bar{f}}(d)S\hat{W}_p(d)/m(d)\} .$$

While one can produce Taylor-Series approximations for the separate stagewise covariance and cross-covariance matrices $\Sigma_P(\mathrm{STAGE})$ and $\Sigma R_P(\mathrm{STAGE})$, where STAGE represents a generic design stage assuming the levels PSU, SCH (school), and STU (student) for the NAEP design, such approximations require the calculation of the PSU and school level double draw probabilities $\phi(\ell\ell')$ and $\phi(ss'|\ell)$. On the other hand, simple analysis of variance type estimators exist for the following composite component matrices

$$S_P(\mathrm{STU}) \equiv \Sigma_P(\mathrm{STU}) - \Sigma R_P(\mathrm{STU})$$
$$S_P(\mathrm{SCH}) \equiv \Sigma_P(\mathrm{SCH}) - \Sigma R_P(\mathrm{SCH}) + \Sigma R_P(\mathrm{STU})$$
$$S_P(\mathrm{PSU}) \equiv \Sigma_P(\mathrm{PSU}) - \Sigma R_P(\mathrm{SCH}) + \Sigma R_P(\mathrm{STU}).$$

These composite component matrices are relatively easy to estimate and provide the necessary ingredients for parameterizing the following design effect version of the P-value covariance matrix model:

$$V_P(\mathrm{DES}) = V_P(\mathrm{SRS})\ \bar{\omega}_P\ [1 + (ncm-1)R_P(0) + (cm-1)R_P(\mathrm{PSU}) + (m-1)R_P(\mathrm{SCH})]$$

where

$$\bar{\omega}_P \equiv V_P(\mathrm{SRS})^{-1}\ \Sigma_P(\mathrm{TOT})/nsm$$
$$R_P(0) \equiv \Sigma_P(\mathrm{TOT})^{-1}\ \Sigma R_P(\mathrm{PSU})$$
$$R_P(\mathrm{PSU}) \equiv \Sigma_P(\mathrm{TOT})^{-1}\ S_P(\mathrm{PSU})$$
$$R_P(\mathrm{SCH}) \equiv \Sigma_P(\mathrm{TOT})^{-1}\ S_P(\mathrm{SCH})\ .$$

The composite component definitions above also lead to the following useful identity

$$\Sigma_P(\mathrm{TOT}) = \Sigma_P(\mathrm{PSU}) + \Sigma_P(\mathrm{SCH}) + \Sigma_P(\mathrm{STU})$$
$$= \Sigma R_P(\mathrm{PSU}) + S_P(\mathrm{PSU}) + S_P(\mathrm{SCH}) + S_P(\mathrm{STU})\ .$$

This identity combined with the consistent estimator for $\Sigma_P(\mathrm{TOT})$ and the Taylor Series ANOVA estimators for $S_P(\mathrm{PSU})$, $S_P(\mathrm{SCH})$, and $S_P(\mathrm{STU})$ provide a consistent estimate for the component $\Sigma R_p(PSU)$ due to primary stratification and without replacement (or PMR) selection; that is,

$$\Sigma\hat{R}_p(PSU) = [\hat{\Sigma}_p(TOT) - \hat{S}_p(PSU) - \hat{S}_p(SCH) - \hat{S}_p(STU)]\ .$$

The $\hat{S}_p(\ \ )$ matrices are estimated using the Taylor-Series linearized single draw variate vectors

$$\hat{\underset{\sim}{z}}(ijk) = \sum_{\ell=1}^{N}\ \sum_{s=1}^{S(\ell)}\ \sum_{t=1}^{M(\ell s)} \alpha_{\ell st}(ijk)\ \hat{\underset{\sim}{z}}(\ell st)$$

where the (dr)-th element of $\hat{\underset{\sim}{z}}(\ell st)$ has the form

$$\begin{aligned}\hat{z}_{dr}(\ell st) &= M(\ell s)\ X_d(\ell st)\ [Y_r(\ell st)-\hat{P}(dr)]/\hat{X}(d)\ \phi(\ell s)\\ &= ncm\ W(\ell st)\ X_d(\ell st)\ [Y_r(\ell st)-\hat{P}(dr)]/\hat{X}(d)\end{aligned}$$

with $\hat{P}(dr)$ and $\hat{X}(d)$ denoting sample estimates for the corresponding population parameters. Recall that $W(\ell st) = M(\ell s)/ncm\ \phi(\ell s)$ is the sample weight for student listing unit SLU(ℓst). In terms of the $\hat{\underset{\sim}{z}}(ijk)$ linearized single draw variate vectors, one computes the following ANOVA type matrix of mean squares and cross-products:

$$MS_p(STU) = \sum_{i=1}^{n}\ \sum_{j=1}^{c}\ \sum_{k=1}^{m}\ [\hat{\underset{\sim}{z}}(ijk)-\hat{\underset{\sim}{z}}(ij_{\cdot})]\ [\hat{\underset{\sim}{z}}(ijk)-\hat{\underset{\sim}{z}}(ij_{\cdot})]^T/nc(m-1)$$

with $\hat{\underset{\sim}{z}}(ij.)$ denoting the school level sample mean vector

$$\hat{\underset{\sim}{z}}(ij_{\cdot}) = \sum_{k=1}^{m}\ \hat{\underset{\sim}{z}}(ijk)/m\ .$$

The corresponding between school within PSU mean square matrix is

$$MS_p(SCH)\ = \sum_{i=1}^{n}\ \sum_{j=1}^{c}\ [\hat{\underset{\sim}{z}}(ij.) - \hat{\underset{\sim}{z}}(i..)][\hat{\underset{\sim}{z}}(ij.) - \hat{\underset{\sim}{z}}(i..)]^T\ /\ n(c-1)$$

with

$$\underset{\sim}{\hat{z}}(i_{\cdot\cdot}) = \sum_{j=1}^{c} \underset{\sim}{\hat{z}}(ij_{\cdot})/c \ .$$

The between PSU mean square matrix is

$$MS_P(PSU) = \sum_{i=1}^{n} \underset{\sim}{\hat{z}}(i_{\cdot\cdot})\underset{\sim}{\hat{z}}(i_{\cdot\cdot})^T/(n-1)$$

noting that the overall mean of the $\underset{\sim}{\hat{z}}(ijk)$ vectors is the null vector.

In terms of the single draw variate vector covariance and cross covariance matrices defined previously, it is not difficult to show that

$$E\ \{ MS_P(STU)\} = S_P(STU) \equiv [\Sigma_P(STU)-\Sigma R_P(STU)]$$

$$E\ \{MS_P(SCH)-MS_P(STU)/m\} = S_P(SCH) \equiv [\Sigma_P(SCH)-\Sigma R_P(SCH)+\Sigma R_P(STU)]$$

and

$$E\ \{MS_P(PSU)-MS_P(SCH)/c\} = S_P(PSU) \equiv [\Sigma_P(PSU)-\Sigma R_P(PSU)+\Sigma R_P(SCH)] \ .$$

The composite component model and associated component estimators for the P-value covariance matrix $V_P(DES)$ have obvious extensions to the transformed P-value case. With $\underset{\sim}{\hat{H}}$ denoting the matrix of partial derivatives of $\underset{\sim}{G}(\underset{\sim}{P})$ with respect to the elements of $\underset{\sim}{P}$ evaluated at $\underset{\sim}{P} = \underset{\sim}{\hat{P}}$, then

$$\hat{\Sigma}_G(TOT) = \underset{\sim}{\hat{H}}\ \hat{\Sigma}_P(TOT)\underset{\sim}{\hat{H}}^T$$
$$\hat{S}_G(STU) = \underset{\sim}{\hat{H}}\ \hat{S}_P(STU)\ \underset{\sim}{\hat{H}}^T$$
$$\hat{S}_G(SCH) = \underset{\sim}{\hat{H}}\ \hat{S}_P(SCH)\ \underset{\sim}{\hat{H}}^T$$
$$\hat{S}_G(PSU) = \underset{\sim}{\hat{H}}\ \hat{S}_P(SCH)\ \underset{\sim}{\hat{H}}^T$$

and

$$\hat{\Sigma R}_G(PSU) = \hat{\Sigma}_G(TOT)-\hat{S}_G(PSU)-\hat{S}_G(SCH)-\hat{S}_G(STU) \ .$$

With the simple random sampling covariance matrix estimator for $\underset{\sim}{G}(\underset{\sim}{\hat{P}})$ depicted by

$$\hat{V}_G(SRS) = \underset{\sim}{\hat{H}}\ \hat{V}_P(SRS)\ \underset{\sim}{\hat{H}}^T \ ,$$

the corresponding unequal weighting matrix $\hat{\bar{w}}_G$ and the composite correlation matrix analogues $\hat{R}_G(0)$, $\hat{R}_G(\text{PSU})$, and $\hat{R}_G(\text{SCH})$ have the same form as the original P-value component matrix estimators with G replacing P in the defining equations.

A straight forward extension of the Taylor Series linearization argument applied to $\underset{\sim}{G}(\hat{\underset{\sim}{P}})$ provides an analogous model for the design based covariance matrix of

$$\begin{aligned}\underset{\sim}{\beta} &= [X^T\hat{V}_G(\text{SRS})^{-1}X]^{-1}X^T\hat{V}_G(\text{SRS})^{-1}\hat{\underset{\sim}{G}} \\ &= \hat{\underset{\sim}{M}}(\text{SRS})\ \hat{\underset{\sim}{G}}\ ,\end{aligned}$$

the SRS based weighted least squares estimate for the $\underset{\sim}{G}(\underset{\sim}{P}) = X\underset{\sim}{\beta}$ linear model coefficients. While one might initially question the use of the SRS based covariance matrix in the definition of $\hat{\underset{\sim}{\beta}}$ above, recall that use of $\hat{V}_G(\text{DES})$ implies full knowledge of the design based P-value covariance matrix $\hat{V}_G(\text{DES})$ which in turn provides for calculation of design based Wald statistics; that is to say, if one use $\hat{V}_G(\text{DES})$ in the definition of $\hat{\beta}$ then no extension of Rao and Scott's approximate methods are required. Returning to our SRS based $\hat{\beta}$, one can further consider a matrix of estimated contrasts $C\hat{\underset{\sim}{\beta}}$. As far as the first order Taylor Series linearization is concerned $C\hat{\underset{\sim}{\beta}}$ is equivalent to a linear transformation of the original vector $\hat{\underset{\sim}{P}}$ of domain P-values; that is

$$C\hat{\underset{\sim}{\beta}} \doteq [C\underset{\sim}{M}(\text{SRS})\ \underset{\sim}{H}]\ \hat{\underset{\sim}{P}}\ .$$

The corresponding Taylor Series component estimators have the form

$$\hat{\Sigma}_{C\beta}(\text{TOT}) = (C\underset{\sim}{\hat{M}\hat{H}})\ \hat{\Sigma}_P(\text{TOT})\ (C\underset{\sim}{\hat{M}\hat{H}})^T$$

$$\hat{S}_{C\beta}(STU) = (C\hat{\underset{\sim}{M}}\hat{\underset{\sim}{H}})\hat{S}_P(STU)\ (C\hat{\underset{\sim}{M}}\hat{\underset{\sim}{H}})^T$$

$$\hat{S}_{C\beta}(SCH) = (C\hat{\underset{\sim}{M}}\hat{\underset{\sim}{H}})S_P(SCH)\ (C\hat{\underset{\sim}{M}}\hat{\underset{\sim}{H}})^T$$

and

$$\hat{S}_{C\beta}(PSU) = (C\hat{\underset{\sim}{M}}\hat{\underset{\sim}{H}})S_P(PSU)\ (C\hat{\underset{\sim}{M}}\hat{\underset{\sim}{H}})^T .$$

where $\hat{\underset{\sim}{M}}$ is shorthand notation for $\hat{\underset{\sim}{M}}(SRS)$. For a NAEP style three-stage design, the generalized design effect matrix for $C\hat{\beta}$ is therefore of the form

$$\begin{aligned} DEFF(\underset{\sim}{C\hat{\beta}}) &= \{C[X^T V_G(SRS)^{-1}X]^{-1}C'\}^{-1}\{(\underset{\sim\sim}{CMH})V_P(DES)(\underset{\sim\sim}{CMH})^T\} \\ &= \bar{\omega}_{C\beta}[I+(ncm-1)\ R_{C\beta}(0)+(cm-1)\ R_{C\beta}(PSU)+(m-1)\ R_{C\beta}(SCH)]. \end{aligned}$$

2.3.5 Asymptotic Distribution of SRS based NAEP Wald Statistics

The partitioning of the generalized design effect matrix for $C\hat{\beta}$ developed in the previous section leads to the following representation for the asymptotic distribution of the SRS based Wald statistic

$$\chi^2_{SRS}(C\hat{\beta}) = \sum_{a=1}^{A} \bar{\omega}_a[1+(ncm-1)\rho_a(0)+(cm-1)\rho_a(PSU)+(m-1)\rho_a(SCH)]\chi^2_a$$

where the χ^2_a are independent single degree of freedom chi-square random variables with coefficients defined in terms of the left and right hand eigen vectors of $DEFF(C\hat{\beta})$, say $\underset{\sim}{L}$ and $\underset{\sim}{R}$, and the component matrices $\bar{\omega}_{C\beta}$, $R_{C\beta}(0)$, $R_{C\beta}(PSU)$, and $R_{C\beta}(SCH)$. Specifically, if $\underset{\sim}{\ell}_a$ is the a-th row of $\underset{\sim}{L}$ and $\underset{\sim}{r}_a$ is the a-th column of $\underset{\sim}{R}$, then the generalized effects of unequal weighting, stratification, PMR selection, and clustering are defined as

$$\bar{\omega}_a = \{ \underset{\sim}{\ell}_a \, \bar{\omega}_{C\beta} \, \underset{\sim}{r}_a \}$$

$$\rho_a(O) = \{ \underset{\sim}{\ell}_a \, [\bar{\omega}_{C\beta} R_{C\beta}(o) \underset{\sim}{r}_a \} \div \bar{\omega}_a$$

$$\rho_a(PSU) = \{ \underset{\sim}{\ell}_a \, [\bar{\omega}_{C\beta} R_{C\beta}(PSU)] \, \underset{\sim}{r}_a \} \div \bar{\omega}_a$$

and

$$\rho_a(SCH) = \{ \underset{\sim}{\ell}_a \, [\bar{\omega}_{C\beta} R_{C\beta}(SCH)] \, \underset{\sim}{r}_a \} \div \bar{\omega}_a \, .$$

For unstratified PPS with replacement selections at each stage of sampling, the $\Sigma R_P(\;)$ cross-variance component matrices are null so that the composite components $S_P(\;)$ equal the corresponding covariance components $\Sigma_P(\;)$. In this case simple single degree of freedom contrasts $C\hat{\underset{\sim}{P}}$ and more complex single degree of freedom contrasts $C\hat{\underset{\sim}{\beta}}$ will have positive design effects of the form

$$DEFF(C\hat{\beta}) = \bar{\omega}_{C\beta} \; [1+(cm-1)\rho_{C\beta}(PSU)+(m-1)\rho_{C\beta}(SCH)]$$

since the cluster correlations

$$\rho_{C\beta}(STAGE) = \{(\underset{\sim}{CMH}) \; \Sigma_P(TOT)(\underset{\sim}{CMH})^T\}^{-1}\{(\underset{\sim}{CMH}) \; \Sigma_P(STAGE)(\underset{\sim}{CMH})^T\}$$

must be nonnegative. This follows from the fact that the $\Sigma_P(TOT)$ and $\Sigma_P(STAGE)$ matrices are all positive definite and

$$\Sigma_P(TOT) = \Sigma_P(PSU) + \Sigma_P(SCH) + \Sigma_P(STU) \, .$$

As indicated earlier, the $\bar{\omega}_P$ matrix for the R = 2 correct-incorrect response pattern case is a D × D diagonal matrix with elements $\bar{\omega}_{C\beta}$ taking the form

$$\bar{\omega}_{C\beta} = \frac{\sum_{d=1}^{D} q_d^2\{P(d)[1-P(d)]/m(d)\} \; \bar{\omega}_P(d)}{\sum_{d=1}^{D} q_d^2 P(d)[1-P(d)]/m(d)}$$

where q_d denotes the d-th element of $\underset{\sim}{CMH}$. Notice that $\bar{\omega}_{C\beta}$ is a weighted average of the $\bar{\omega}_P(d)$ quantities with the weights $q_d^2P(d)[1-P(d)]/m(d)$ all positive and the $\bar{\omega}_p(d)$ all expected to exceed 1. Therefore, one should expect $\bar{\omega}_{C\beta}$ to exceed 1.

While one should therefore expect design effects for single degree of freedom P-value contrasts to exceed 1 for unstratified with replacement cluster samples, the tendency for the PSU and SCH (school) stage specific covariance matrices $\Sigma_p(\quad)$ to have positive covariance terms for domains d and d' that are typically represented in the same schools and PSUs will cause contrasts among such domain P-values to have smaller cluster correlations than observed for the separate domain P-values. Lepkowski and Landis (1980) examining data from the Health Examination Survey (HES) and a 1974 University of Michigan Survey Research Center Omnibus (OMNI) Survey, observed this tendency for the DEFF of P-value contrasts to be substantially smaller than the DEFF of individual P-values. The size of the proportional reduction in an average contrast DEFF relative to the average P-value DEFF was 60 percent for the HES and 9 percent for the OMNI survey. The size of the proportional reduction factors observed by Lepkowski and Landis depended on the magnitude of the average P-value DEFF. For HES where the P-value DEFF's averaged 3.91, an overall 60 percent reduction was observed. For the OMNI survey where the P-value DEFF's averaged 1.10, the overall proportional reduction for contrast DEFF's was only 9 percent. A similar tendency for the proportional reduction to vary with the mean P-value DEFF was observed across dependent variates within the two surveys.

For designs with stratification and without-replacement or PMR selections at the various stages, the general expressions for the composite components contain cross-covariance matrices $\Sigma R_p(PSU)$ and $\Sigma R_p(SCH)$ that are expected to be negative definite. In this general case, the PSU and SCH correlation coefficients

$$\rho_{C\beta}(PSU) = \{\underset{\sim}{q}^T\Sigma_p(PSU)[I-R_p(PSU)]\underset{\sim}{q} + \underset{\sim}{q}^T\Sigma R_p(SCH)\underset{\sim}{q}\} \div \{\underset{\sim}{q}^T\Sigma_p(TOT)\underset{\sim}{q}\}$$

$$= [\delta_{C\beta}(PSU) - \zeta_{C\beta}(SCH)]$$

and

$$\rho_{C\beta}(SCH) = \{\underset{\sim}{q}^T\Sigma_p(SCH)[1-R_p(SCH)]\underset{\sim}{q} + \underset{\sim}{q}^T\Sigma R_p(STU)\underset{\sim}{q}\} \div \{\underset{\sim}{q}^T\Sigma_p(TOT)\underset{\sim}{q}\}$$

$$= [\delta_{C\beta}(SCH) - \zeta_{C\beta}(STU)]$$

may be negative if the combined stratification and without replacement or PMR selection effects $\zeta_{C\beta}$ from the subsequent stage swamp the clustering effects $\delta_{C\beta}$. The general case also has the primary stage stratification and without replacement or PMR selection effect

$$\rho_{C\beta}(0) = \{\underset{\sim}{q}^T\ \Sigma R_p(PSU)\ \underset{\sim}{q}\}/\{\underset{\sim}{q}^T\Sigma_p(TOT)\underset{\sim}{q}\}$$

$$= -\zeta_{C\beta}(PSU)$$

which is expected to be negative and which has a large coefficient (nsm-1) in the design effect expression.

The empirical results in chapter 3 of this report show that for simple constrasts among NAEP P-values and for weighted-least squares coefficients, a substantial fraction of the design effects are less than 1. Lepkowski and Landis also observed numerous contrast DEFFs less than 1. In fact, the OMNI data had mean contrast DEFFs for the ten dependent variables they explored ranging from 0.75 to 1.19 with an average of 0.99. In such instances, the SRS based chi-squared statistics are smaller and less significant than the design based chi-square. While one might attribute some of these DEFF values less than 1 to negative bias in the design based Taylor Series variance approximation, we feel that the incidence of such cases is too great to be totally explained in this fashion. Furthermore, the variance formula used with the Taylor Series linearization to produce the estimates of $\hat{V}_p(DES)$ used in this report would overestimate the variance of a linear statistic since it assumes that primary units were selected two or three per stratum with replacement, when in fact they were selected without replacement. Specifically, the linearized single draw variates are formed separately by primary stratum h such that

$$\tilde{z}_{dr}(hijk) = M(hij)\ X_d(hij)\ [Y_r(hij) - \hat{P}(dr)]/\hat{X}(d)\ \phi(hij)$$

where $\phi(hij) = \phi(hi)\ \phi(j|hi)$ denotes the nonresponse adjusted single draw probability for sample school j of sample PSU(i) based on n(h) = 2 or 3 PSU selections from primary stratum h and c(hi) school selections from sample PSU(hi). The PSU level averages

$$\hat{\underset{\sim}{z}}(hi_{..}) = \sum_{j=1}^{c(hi)} \sum_{k=1}^{m(hij)} \underset{\sim}{z}(hijk)/c(hi)\ m(hij)$$

are then formed with m(hij) denoting the number of package respondents from sample school (hij). The P-value covariance matrix is then estimated by the between PSU within stratum mean square

$$\hat{V}_p(DES) = \sum_{h=1}^{H} \sum_{i=1}^{n(h)} [\hat{\underset{\sim}{z}}(hi_{..})-\hat{\underset{\sim}{z}}(h_{...})]\ [\hat{\underset{\sim}{z}}(hi_{..})-\hat{\underset{\sim}{z}}(h_{...})]^T/n(h)[n(h)-1]\ .$$

$$= \sum_{h=1}^{H} MS_p(PSU|h)/n(h)$$

where

$$\hat{\underset{\sim}{z}}(h_{...}) = \sum_{i=1}^{n(h)} \hat{\underset{\sim}{z}}(hi_{..})/n(h)$$

is the primary stratum-h mean of the linearized single draw variate vector. If the within stratum h primary selections had been with replacement, then the variance estimator above would be unbiased for a linear statistic like the vector of domain by response option totals $\hat{\underset{\sim}{Y}}(+)$. Since the NAEP primaries were selected without replacement, the stratum h effect of without replacement selection, namely $\Sigma R_y(PSU|h)$, is not accounted for. This matrix is expected to be negative definite so that its exclusion from the variance function will enlarge the variance of any contrast.

In the Year 13 NAEP primary sample where sequential PMR selections were made from a judicously ordered primary frame, the pseudo-strata formed by pairing neighboring selections down the ordered listing should also lead to some positive bias in the variance approximation due to ignoring the deeper implicit stratification. To explore this issue further one could contrast Wald statistics based on the Taylor Series covariance matrix estimator with Wald statistics derived from Balanced Repeated Replication (BRR) covariance matrix estimators. Krewski and Rao's (1979) small sample comparisons of TSL and BRR variances for combined ratio estimators suggests that TSL generally has a negative bias while BRR has a positive bias under the model

$$Y(hi) = \alpha(h) + \beta(h)X(hi) + e(hi)$$

with

$$E\,[e(hi)|X(hi)] = 0$$

and

$$E\,[e^2(hi)|X(hi)] = \delta_h X(hi)^t\ .$$

Rao and Krewski show that the absolute bias comparison favors TLS when $t \leq 1$. When $t = 2$, the BRR variance estimators have smaller absolute bias. In terms of mean-squared error, the results of Krewski and Rao (1979) and Frankel (1971) suggest that the TSL variances are generally more accurate estimators. On the other hand, Frankel (1971) and Campbell and Meyer (1978) show that in reasonably small samples BRR may produce more robust inferences in terms of achieving the desired significance level. Direct TSL and BRR comparisons would shed some light on the TSL negative bias potential. Unfortunately such comparisons were beyond the scope of the current project.

2.4 Testing Balanced Fits Via Dummy Variable Regression

An alternative mode of analysis for exploring the effect of domain classifiers on the $Y_r(\ell st)$ zero-one correct response indicators has been referred to as "Balanced Fitting" by NAEP analysts. This approach utilizes dummy variable regression models of the form

$$\hat{Y}_r(\ell st) = \underset{\sim}{X}(\ell st)\ \underset{\sim}{B}\ ,$$

where the row vector $\underset{\sim}{X}$ of independent variables includes a leading 1 for the intercept parameter and zero-one indicator variables for parameters associated with the levels of student's race, sex, parent's education, and type of community where the school is located. The proper sample design based analysis for testing the significance of such regression coefficients has been specified by Folsom (1974). The universe level least-squares solution for the vector of regression coefficients $\underset{\sim}{B}$ is specified in terms of the universe level left and right hand sides of the so-called normal equations; namely

$$(X^T X) = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} \underset{\sim}{X}(\ell st)^T \underset{\sim}{X}(\ell st)$$

and

$$(X^T Y) = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} \underset{\sim}{X}(\ell st)^T Y(\ell st)\ .$$

With the regression model parameterized such that $(X^T X)$ is nonsingular, the $\underset{\sim}{B}$ vector is defined as

$$\underset{\sim}{B} = (X^T X)^{-1}(X^T Y)$$

with ($)^{-1}$ denoting matrix inversion.

In terms of the balanced three stage analogue of the NAEP design explored in the previous sections, the unbiased sample estimators for the left and right hand sides are formed from the following single draw variates

$$(x^Tx)_{ijk} = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} \alpha_{\ell st}(ijk)M(\ell s)\underset{\sim}{X}(\ell st)^T\underset{\sim}{X}(\ell st)/\phi(\ell s)$$

and

$$(x^Ty)_{ijk} = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} \alpha_{\ell st}(ijk)M(\ell s)\underset{\sim}{X}(\ell st)^TY(\ell st)/\phi(\ell s) .$$

The corresponding unbiased estimators are formed as the sample means

$$(x^Tx)_{...} = \sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} (x^Tx)_{ijk}/ncm$$

and

$$(x^Ty)_{...} = \sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} (x^Ty)_{ijk}/ncm .$$

The associated sample estimator for the vector of regression coefficients is

$$\hat{\underset{\sim}{B}} = (x^Tx)^{-1}_{...}(x^Ty)_{...} .$$

To approximate the sampling variance of $\hat{\underset{\sim}{B}}$, the following Taylor series linearized variate was derived independently by Folsom (1974) and Fuller (1974):

$$\underset{\sim}{Q}(\ell st) = (X^TX)^{-1}\ \underset{\sim}{X}(\ell st)^T\ e(\ell st)$$

where

$$e(\ell st) = [Y(\ell st) - \underset{\sim}{X}(\ell st)\underset{\sim}{B}]$$

denotes the prediction error or deviation from regression for student list unit (ℓst). The corresponding list unit single draw variate vector is

$$\underset{\sim}{q}(\ell st) = M(\ell s)\underset{\sim}{Q}(\ell st)/\phi(\ell s) \ .$$

Substituting these linearized single draw variate vectors for the $\underset{\sim}{z}(\ell st)$ vectors used previously to define the $\Sigma_P(\text{STAGE})$ covariance and $\Sigma R_P(\text{STAGE})$ cross-covariance component matrices for the vector $\hat{\underset{\sim}{P}}$ of estimated domain P-values, one obtains an analogous set of $\Sigma_B(\text{STAGE})$ covariance and $\Sigma R_B(\text{STAGE})$ cross-variance components. Recalling the general form for the total covariance matrix $\Sigma_B(\text{TOT})$, one can show that

$$\Sigma_B(\text{TOT}) = M(++)(X^TX)^{-1} \left\{ \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \zeta(\ell s) \sum_{t=1}^{M(\ell s)} \underset{\sim}{X}(\ell st)^T \underset{\sim}{X}(\ell st) e(\ell st)^2 \right\} (X^TX)^{-1}$$

where

$$\zeta(\ell s) = \{M(\ell s)/M(++)] \div \phi(\ell s) \ .$$

This result derives from the fact that

$$\underset{\sim}{q}(\ldots) = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \phi(\ell s) \sum_{t=1}^{M(\ell s)} q(\ell st)/M(\ell s) = \underset{\sim}{\Phi} \ ,$$

with $\underset{\sim}{\Phi}$ denoting the null vector. For a self-weighting sample with $\zeta(\ell s) = 1$ for all (ℓs), the total covariance component matrix $\Sigma_B(\text{TOT})$ is equivalent to the simple random sampling covariance matrix

$$\Sigma_B(\text{SRS}) = M(++)(X^TX)^{-1} \left\{ \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} \underset{\sim}{X}(\ell st)^T \underset{\sim}{X}(\ell st) e(\ell st)^2 \right\} (X^TX)^{-1} \ .$$

The estimated linearized variate for $\underset{\sim}{B}$ is defined as follows

$$\hat{\underset{\sim}{q}}(\ell st) = M(\ell s) \ (x^Tx)^{-1}_{\cdots} \ \underset{\sim}{X}(\ell st)^T r(\ell st)/\phi(\ell s)$$

with

$$r(\ell st) = [Y(\ell st) - \underset{\sim}{X}(\ell st) \ \hat{\underset{\sim}{B}}]$$

denoting the observed sample residuals. Composite component matrices $S_B(\text{PSU})$, $S_P(\text{SCH})$, and $S_B(\text{STU})$ are estimated from the analogous ANOVA type matrix mean squares $MS_B(PSU)$, $MS_B(SCH)$, and $MS_B(STU)$. These $\underset{\sim}{\hat{B}}$ mean-square matrices are defined by analogy with the corresponding MS_P matrices using

$$\underset{\sim}{\hat{q}}(ijk) = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} \alpha_{\ell st}(ijk)\underset{\sim}{\hat{q}}(\ell st)$$

in place of the $\underset{\sim}{\hat{z}}(ijk)$ linearized single draw variate vectors.

A consistent estimator for $\Sigma_B(TOT)$ is obtained by recasting $\Sigma_B(TOT)$ in a form involving the sample weights $W(\ell st) = M(\ell s)/ncm\ \phi(\ell s)$; that is,

$$\Sigma_B(TOT) = ncm \left\{ \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} W(\ell st)\underset{\sim}{q}(\ell st)\underset{\sim}{q}(\ell st)^T \right\}$$

$$= f\bar{\omega}\ M^2(++) \left\{ \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} W(\ell st)\underset{\sim}{Q}(\ell st)\underset{\sim}{Q}(\ell st)^T/W(+++) \right\}$$

where

$$W(+++) = \sum_{\ell=1}^{N} \sum_{s=1}^{S(\ell)} \sum_{t=1}^{M(\ell s)} W(\ell st)$$

is the universe weight sum;

$$\bar{\omega} = W(+++)/M(++)$$

is the universe level average weight, and

$$f = ncm/M(++)$$

is the overall sampling fraction. As before, the unequal weighting design effect is estimated by

$$\widehat{f\bar{\omega}} = ncm \sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} W(ijk)^2/\left[\sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} W(ijk)\right]^2 .$$

The matrix inside of curley brackets, say $SW_B(TOT)$ is estimated consistently by

$$\hat{SW}_B(TOT) = \sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} W(ijk)^2\underset{\sim}{\hat{Q}}(ijk)\underset{\sim}{\hat{Q}}(ijk)^T/ \sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} W(ijk)^2$$

where

$$\hat{\underset{\sim}{Q}}(ijk) = (x^T x)^{-1}_{\cdots}\ \underset{\sim}{X}^T(ijk)r(ijk)$$

is the estimated Taylor Series linearized variate without the division by our single draw probability $\phi(\ell st) = \phi(\ell s)/M(\ell s)$.

The simple random sampling covariance matrix $\Sigma_B(SRS)/M(++)^2$ is similarly approximated by

$$\hat{S}_B(SRS) = \sum_{i=1}^{n}\sum_{j=1}^{c}\sum_{k=1}^{m} W(ijk)\hat{\underset{\sim}{Q}}(ijk)\hat{\underset{\sim}{Q}}(ijk)^T / \sum_{i=1}^{n}\sum_{j=1}^{c}\sum_{k=1}^{m} W(ijk)\ .$$

Notice that $\widehat{SW}_B(TOT)$ is the weighted sample mean of the $\hat{\underset{\sim}{Q}}(ijk)\hat{\underset{\sim}{Q}}(ijk)^T$ matrices using squared weights $W(ijk)^2$ while $S_B(SRS)$ is the comparable weighted average based on the original sample weights. For the statistic $\hat{\underset{\sim}{B}}$, the effect of unequal weighting

$$\bar{\omega}_B = \Sigma_B(SRS)^{-1}\ \Sigma_B(TOT)$$

is estimated by

$$\hat{\bar{w}}_B = \hat{f}\hat{\bar{w}}\hat{S}_B(SRS)^{-1}\ SW_B(TOT)\ .$$

The generalized design effect matrix for $\hat{\underset{\sim}{B}}$ has another component that arises from the typical model based least-squares analysis. Assuming that

$$[W(ijk)]^{\frac{1}{2}}Y(ijk) = [W(ijk)]^{\frac{1}{2}}\ \underset{\sim}{X}(ijk)\underset{\sim}{B} + e(ijk)$$

with errors having zero expectation and common variance σ_e^2 conditional on the given set of $\underset{\sim}{X}(ijk)$ and $W(ijk)$ variables, ordinary least-squares theory produces our weighted $\hat{\underset{\sim}{B}}$ coefficients, and the model based covariance matrix

$$\hat{V}_B(MOD) = (x^T x)^{-1}_{\cdots}\ \hat{\sigma}_e^2$$

where

$$\hat{\sigma}_e^2 = \sum_{i=1}^{n} \sum_{j=1}^{c} \sum_{k=1}^{m} W(ijk)r(ijk)^2/(ncm - p)$$

is the residual mean square of the $W^{\frac{1}{2}}$ transformed variables. Recognizing that

$$nsm\ \hat{V}_B(SRS) = \hat{M}(++)\ (x^Tx)^{-1}_{\dots}\{\sum_{ijk} W(ijk)\underset{\sim}{X}^T(ijk)\underset{\sim}{X}(ijk)r(ijk)^2\}\ (x^Tx)^{-1}_{\dots}$$

with the total student population size M(++) estimated by

$$\hat{M}(++) = \sum_{ijk} W(ijk),$$

one can write the estimated model effect as

$$\begin{aligned}\hat{M}_B &= \hat{V}_B(MOD)^{-1}\hat{V}_B(SRS)\\ &= [(ncm-p)/ncm]\{\sum_{ijk} W(ijk)\underset{\sim}{X}^T(ijk)\underset{\sim}{X}(ijk)r(ijk)^2/\sigma w_e^2\}(x^Tx)^{-1}_{\dots}\end{aligned}$$

where

$$\sigma w_e^2 = \sum_{ijk} W(ijk)r(ijk)^2/\sum_{ijk} W(ijk)$$

is the weighted residual mean square. Under the model where $E\{r(ijk)^2\} \doteq \sigma_e^2$ for large samples, the model expectation of $\hat{M}_B$ given the sample will be approximately 1.

For a set of linear contrasts among the $\underset{\sim}{B}$ coefficients, say $C\underset{\sim}{B}$, one can define corresponding component matrices

$$\begin{aligned}S_{CB}(STAGE) &= C\ S_B(STAGE)C^T\\ \Sigma_{CB}(TOT) &= C\ \Sigma_B(TOT)C^T\\ \Sigma_{CB}(SRS) &= C\ \Sigma_B(SRS)C^T\end{aligned}$$

and

$$V_{CB}(MOD) = C\ V_B(MOD)C^T\ .$$

These components lead to an estimated generalized design effect matrix of the form

$$\mathrm{DEFF}(\underset{\sim}{CB}) = \hat{M}_{CB}\hat{\bar{\omega}}_{CB}[I+(ncm-1)\hat{R}_{CB}(0)+(cm-1)\hat{R}_{CB}(PSU)+(m-1)\hat{R}_{CB}(SCH)]$$

where

$$\hat{R}_{CB}(0) = [I-\hat{R}_{CB}(PSU)-\hat{R}_{CB}(SCH)-\hat{R}_{CB}(STU)]$$

and

$$\hat{R}_{CB}(STAGE) = \hat{\Sigma}_{CB}(TOT)^{-1}\hat{S}_{SB}(STAGE) .$$

With these results, one can again write the asymptotic distribution of the ordinary least squares model based test statistic as

$$\chi^2_{MOD}(\widehat{CB}) \sim \sum_{a=1}^{A} M_{CB}(a)\bar{\omega}_{CB}(a)[1+(ncm-1)\rho^a_{CB}(0)+(cm-1)\rho^a_{CB}(PSU)+(m-1)\rho^a_{CB}(SCH)]\chi^2_a$$

where the χ^2_a are independent single degree of freedom central chi-square variables and the coefficient components are of the form

$$M_{CB}(a) = \{\underset{\sim}{\ell}_a M_{CB}\underset{\sim}{r}_a\}$$

$$\bar{\omega}_{CB}(a) = \{\underset{\sim}{\ell}_a M_{\underset{\sim}{CB}}\bar{\omega}_{\underset{\sim}{CB}}\underset{\sim}{r}_a\} \div M_{CB}(a)$$

$$\rho^a_{CB}(STAGE) = \{\underset{\sim}{\ell}_a M_{CB}\bar{\omega}_{CB}R_{CB}(STAGE)\underset{\sim}{r}_a\} \div M_{CB}(a)\bar{\omega}_{CB}(a)$$

where $\underset{\sim}{\ell}_a$ is the a-th row of the left hand eigenvectors of DEFF($\underset{\sim}{CB}$) and $\underset{\sim}{r}_a$ is the a-th column of the corresponding right-hand eigenvectors.

The empirical results presented in chapter 3 for NAEP balanced fit parameters suggests that the design effects for these statistics are generally greater than one. We suspect that the extra model effect components $M_{CB}(a)$ contribute substantially to this result.

2.5 Inference for NAEP Package Means

In order to increase the precision of subgroup comparisons, NAEP analysts have turned to averages of single exercise P-values. Averaging across exercises (e) within packages (indexed by u) should reduce the variance inflating effect of stochastic response errors. Averaging exercises across packages has the potential to substantially reduce sampling errors, since each distinct package contributes a nonoverlapping sample of approximately 2,600 students. Recalling the definition of an estimated NAEP P-value presented in section 2.1, a within package average across exercises labeled e = 1, 2, . . . E(u) can be written in terms of the weighted mean of a student level proportion correct variable. With $Y_{ue}(hijk)$ denoting the correct incorrect response indicator for exercise e of package u from sample school j of PSU(i) as administered to sample student (hijk) in primary stratum (h), the weighted package u mean for domain d is

$$\hat{P}_{u\cdot}(d) = \sum_{e=1}^{E(u)} \hat{P}_{ue}(d)/E(u)$$

$$= \{ \sum_{h=1}^{H} \sum_{i=1}^{n(h)} \sum_{j=1}^{c_u(hi)} \sum_{k=1}^{m_u(hij)} W_u(hijk) X_{du}(hijk) [\sum_{e=1}^{E(u)} Y_{ue}(hijk)/E(U)] \} / \hat{X}_u(d)$$

where $m_u(hij)$ denotes the number of package u respondents from the j-th cooperating package u school from PSU(hi) with j ranging over $c_u(hi)$ such schools. The denominator of $\hat{P}_u(d)$ is the package u weight sum for domain d members.

To estimate the design based covariance matrix for the vector $\hat{\underset{\sim}{P}}(u)^T = [\hat{P}_{u\cdot}(1), \ldots, \hat{P}_{u\cdot}(d), \ldots, \hat{P}_{u\cdot}(D)]$ of package u means, the student level P-values

$$Y_{u\cdot}(hijk) = \sum_{e=1}^{E(u)} Y_{ue}(hijk)/E(u)$$

are used to form linearized variates

$$\hat{\mu}_{du}(hijk) = M_u(hij)X_{du}(hijk)[Y_{u.}(hijk)-\hat{P}_{u.}(d)]/\hat{X}_u(d)\phi_u(hij)$$
$$= n(h)c_u(hi)m_u(hij)W_u(hij)X_{du}(hijk)[Y_{u.}(hijk)-\hat{P}_{u.}(d)]/\hat{X}_u(d)$$

These linearized variates lead to the following Taylor Series covariance matrix estimator based on the paired with replacement PSU selection model

$$\hat{V}^{u}_{\bar{P}}(DES) = \sum_{h=1}^{H} MS^{u}_{\bar{P}}(h)/n(h)$$

where

$$MS^{u}_{\bar{P}}(h) = \sum_{i=1}^{n(h)} [\underset{\sim}{\hat{\mu}}_u(hi..)-\underset{\sim}{\hat{\mu}}_u(h...)][\underset{\sim}{\hat{\mu}}_u(hi..)-\underset{\sim}{\hat{\mu}}_u(h...)]^T/[n(h)-1],$$

is the primary stratum h contribution to the covariance matrix. To produce the PSU(hi) level mean vectors $\underset{\sim}{\hat{\mu}}_u(hi..)$, the student level vectors of D linearized variates

$$\underset{\sim}{\hat{\mu}}_u(hijk)^T = [\hat{\mu}_{1u}(hijk), \ldots, \hat{\mu}_{du}(hijk), \ldots, \hat{\mu}_{Du}(hijk)] ,$$

are first averaged over the $m_u(hij)$ students responding to package u in cooperating school (hij) and then these school level mean vectors are averaged over the $c_u(hi)$ cooperating schools from PSU(hi) that are assigned package u.

To allow for the consideration of item P-value averages extending across packages, say

$$\hat{P}_{..}(d) = \sum_{u=1}^{U} \hat{P}_{u.}(d)/U ,$$

the full covariance matrix for the extended vector

$$\underset{\sim}{\hat{P}}^T = [\underset{\sim}{\hat{P}}(1)^T, \ldots, \underset{\sim}{\hat{P}}(u)^T, \ldots, \underset{\sim}{\hat{P}}(U)]$$

of package level domain means is required. This extended covariance matrix can be produced simply by extending the PSU level linearized vector means to include subvectors for each package u involved in the average; that is, one defines

$$\hat{\underset{\sim}{U}}(hi_{..})^T = [\hat{\underset{\sim}{\mu}}_1(hi_{..})^T, \ldots, \hat{\underset{\sim}{\mu}}_u{}^T(hi_{..}), \ldots, \hat{\underset{\sim}{\mu}}_U(hi_{..})^T]$$

and forms $\hat{V}_{\underset{\sim}{P}}$(DES)by substituting $\hat{\underset{\sim}{U}}(hi_{..})$ and $\hat{\underset{\sim}{U}}(h_{...})$ for the associated package specific vectors in the definition of $\hat{V}^{u}_{\underset{\sim}{P}}$(DES). Note that the vector of D cross-package means is a simple linear transformation of the $\hat{\underset{\sim}{P}}$ vector of the form

$$\hat{\bar{\underset{\sim}{P}}} = C\hat{\underset{\sim}{P}}$$

with the d-th row of C having the form

$$\underset{\sim}{C}(d) = [\underset{\sim}{\delta}_1(d), \ldots, \underset{\sim}{\delta}_u(d), \ldots, \underset{\sim}{\delta}_u(d)]/U$$

where $\underset{\sim}{\delta}_u(d)$ is a (1 × D) row vector with a 1 in position d and zeros elsewhere. The estimated design based covariance matrix for $\hat{\bar{\underset{\sim}{P}}}$ is therefore

$$\hat{V}_{\bar{P}}(DES) = C\hat{V}_{\underset{\sim}{P}}(DES)C^T .$$

The simple random sampling covariance matrix $\hat{V}_{\bar{P}}$(SRS) for $\hat{\bar{\underset{\sim}{P}}}$ is diagonal with d-th diagonal element

$$\hat{V}^{d}_{\bar{P}}(SRS) = \{\sum_{u=1}^{U} S^2_u(d)/m_u(d)\}/U^2$$

where

$$S^2_u(d) = \sum_{h=1}^{H} \sum_{i=1}^{n(h)} \sum_{j=1}^{c_u(hi)} \sum_{k=1}^{m_u(hij)} W_u(hijk)X_{du}(hijk)[Y_{u.}(hijk)-P_{u.}(d)]^2/\hat{X}_u(d)$$

is the estimated subpopulation d variance of the $Y_{u.}(hijk)$ student level proportions correct. Recall that $\hat{X}_u(d)$ is the package u estimate of the universe count of students in subpopulation d.

63

The chi-square adjustment factors proposed by Rao and Scott based on the average eigenvalue of the generalized design effect matrix

$$\text{DEFF}[\hat{\bar{\underset{\sim}{P}}}] = \hat{V}_{\bar{P}}(\text{SRS})^{-1}\ \hat{V}_{\bar{P}}(\text{DES})$$

have the form

$$\text{AVED}[\bar{\underset{\sim}{P}}] = \sum_{d=1}^{D} \text{DEFF}[\hat{\bar{P}}(d)]/D$$

were DEFF[$\hat{\bar{P}}(d)$] is the design effect for the d-th element of $\hat{\bar{\underset{\sim}{P}}}$.

In addition to the weighted least squares/Wald statistic type analysis directed at the $\hat{\bar{\underset{\sim}{P}}}$ vector, balanced fit type analyses directed at cross-exercise and cross-package means have been pursued. For these analyses, the student level P-values $Y_{u.}(hijk)$ for all the U package samples involved in the cross-package average were used as the dependent variables in a pair of main effect regression models. For a fully interactive regression model including for example race, sex, and parents education, the model based predicted values for each race by sex by parents education cell (c) would have the form of a weighted combined ratio mean

$$\hat{\bar{Y}}(c) = \left\{ \frac{\sum\limits_{u=1}^{U} \sum\limits_{hijk} W_u(hijk) X_{uc}(hijk) Y_{u.}(hijk)}{\sum\limits_{u=1}^{U} \sum\limits_{hijk} W_u(hijk) X_{uc}(hijk)} \right\}$$

$$= \left\{ \sum_{u=1}^{U} \hat{X}_u(c) \hat{P}_{u.}(c) / \sum_{u=1}^{U} \hat{X}_u(c) \right\}$$

where $\hat{X}_u(c)$ is the package u sample estimate of the universe level student count for subpopulation c. The main effect balanced fit models yield reduced model approximations of the combined ratio means $\hat{\bar{Y}}(c)$. The (DES) covariance matrix for the main effect parameters fit to these student level P-values were obtained using the design based regression procedures described in section 2.4. The corresponding covariance matrix applicable for a standard model based regression analysis were obtained by running the transformed variables $W_u(hijk)^{\frac{1}{2}} Y_{u\cdot}(hijk)$ and $W_u(hijk)^{\frac{1}{2}} \underset{\sim}{X}_u(hijk)$ through an ordinary least squares package yielding

$$\hat{V}_{\bar{\beta}}(MOD) = (x^T x)_{....}^{-1} \hat{\sigma}_e^2$$

where $\hat{\sigma}_e^2$ is the residual mean square among the transformed Y variates and $(x^T x)^{-1}_{....}$ denotes the inverse of the weighted sums of squares and cross products matrix forming the left-hand sides of the normal equations.

3. EMPIRICAL INVESTIGATION

3.1 Analysis Items and Subgroups

Intially, five NAEP exercises per age class were selected for analysis from the Year 09 Mathematics Assessment. One item was selected from each of the following five content objectives:

A. numbers and numeration,
B. variables and relationships,
C. size, shape, and position,
D. measurement, and
E. other topics.

Copies of the selected exercises are included in Appendix A. Each item was recoded one for correct and zero for incorrect. An additional score was defined for each student as the proportion of the items analyzed on a package that the student answered correctly. This score was analyzed within each age class to form three mean scores for analysis.

Four domain or subgroup defining variables were also selected. These were, with their corresponding levels:

Sex
- Male
- Female

Race
- White
- Other

Type of Community (TOC)
- Extreme Rural
- Metro
- Other

Parental Education (PARED)
- Not High School Graduate
- High School Graduate
- Post High School

3.2 Analyses

The ultimate goal of this study was to compare sample design based analyses of NAEP data with those assuming a simple random sample. This was done separately for two analytic methods. The first analytic method that will be discussed is the Wald statistic/weighted least squares approach. This will be followed by a discussion of the work done for the balanced fits analyses.

The Wald statistic/weighted least squares approach, described earlier in section 2.1, proceeds by first estimating a vector of domain statistics and its corresponding covariance matrix. Various hypotheses concerning this vector can then be evaluated. Two vectors of domain means were formed for each of the 15 item scores and the three mean scores. The first vector contained 12 elements corresponding to the complete cross-classification of Race, Sex and Parents Education (PARED). The second vector was derived from the cross-classification of Sex, Type of Community (TOC) and PARED and was of length 18. For the 15 item scores, these vectors consisted of simple proportion correct p-values. Two covariance matrices were then estimated for each vector. One based upon the actual sample design and the other assuming a simple random sample of students. The details of the estimation process were provided in Chapter 2.

At this point several exercises were excluded from the study because their estimated covariance matrices were singular. For the Race*Sex*PARED cross-classification only item N0317A was excluded. However, for the Sex*TOC*PARED cross-classification it was necessary to exclude items N0227A, N0317A, N0323A, T0224A, and S0121A.

A linear model was then fitted, via weighted least squares, to each of the remaining domain mean vectors. For the Race*Sex*PARED domain cross-classification vectors the model contained the main effects of Race and Sex, a linear effect of PARED and the four possible two- and three-way interactions among these three effects. The Sex*TOC*PARED domain classification model had the same form except that TOC was substituted for Race. These models were fitted two ways -- one weighted with the design based covariance matrix and the other weighted with the simple random sampling covariance matrix. The lack of fit of each model and the significance of each effect in the model was then assessed. These tests are labelled one through eight in Tables 3-1 and 3-2.

In addition, nine other hypotheses were considered and are labeled nine through 17 in Tables 3-1 and 3-2. These hypotheses were tested via direct contrasts of the domain means. The tests labeled "average" (numbers 10, 11, 12 and 13) average the effect over the combined levels of the other two variables. On the other hand, the "nested" tests (numbers 14, 15, 16 and 17) test for all the indicated simple effects being simultaneously null over the combined levels of the other two variables.

Three test statistics were entertained for each hypothesis. The first test was a Wald statistic chi-squared based upon the actual NAEP sample design. A second Wald statistic chi-squared was also calculated assuming a simple random sample of students. Finally, the simple random sampling chi-squared was adjusted as shown in section 2.2 by dividing by the average design effect to obtain the third test statistics. These three test statistics were calculated for each hypothesis for 14 NAEP items and three mean scores for the Race*Sex*PARED cross-classification, as well as for 10 NAEP items plus three mean scores for the Sex*TOC*PARED cross-classification. All of these test statistics are shown in Appendix B along with their associated significance levels assuming that each has a chi-squared distribution. The test numbers in Appendix B correspond to those in Tables 3-1 and 3-2.

Turning now to the balanced effects analyses, the 15 NAEP items plus three age related mean scores discussed earlier were studied. As noted in Chapter 2, the balanced effects methodology is used in a regression setting to assess the significance of a particular effect after adjusting for the other factors in the model. For this portion of the study, each of the NAEP item scores and three mean scores were regressed on two models. One

Table 3-1. Hypothesis Tests for the Race*Sex*PARED Cross-Classification

Test Number	d.f.	Description
Linear Model Tests		
1	4	Lack of fit
2	1	Race
3	1	Sex
4	1	PARED linear
5	1	Race*Sex
6	1	Race*PARED linear
7	1	Sex*PARED linear
8	1	Race*Sex*PARED linear
Contrast Tests		
9	11	All cells equal
10	1	Average Race effect
11	1	Average Sex effect
12	2	Average PARED effect
13	1	Average PARED linear effect
14	6	Nested Race effect
15	6	Nested Sex effect
16	8	Nested PARED effect
17	4	Nested PARED linear effect

69

Table 3-2. Hypothesis Tests for the Sex*TOC*PARED Cross-Classification

Test Number	d.f.	Description
Linear Model Tests		
1	6	Lack of fit
2	1	Race
3	2	TOC
4	1	PARED linear
5	2	Sex*TOC
6	1	Sex*PARED linear
7	2	TOC*PARED linear
8	2	Sex*TOC*PARED linear
Contrast Tests		
9	17	All cells equal
10	1	Average Sex effect
11	2	Average TOC effect
12	2	Average PARED effect
13	1	Average PARED linear effect
14	9	Nested Sex effect
15	12	Nested TOC effect
16	12	Nested PARED effect
17	6	Nested PARED linear effect

model contained the main effect of Sex, Race and PARED, while the other contained the main effect of Sex, TOC and PARED. The three partial F-tests for each effect in the model controlling for the other two effects were then considered for each model and mean or item score.

Each model, and hence each F-test, was fitted in three different ways for comparison. One approach employed the sampling weights and the Taylor series variance estimation technique discussed in section 2.4. This yielded strict design based significance tests. Test statistics were also obtained using a standard regression package (the GLM procedure of SAS) ignoring both the sample design and the sampling weights. This approach produces biased estimates of the regression coefficient, as well as producing inferential statistics under inappropriate standard regression assumptions. Finally, a weighted version of the SAS GLM procedure was used. This process properly incorporates the sampling weights to produce the correct statistically consistent estimates of the regression coefficients while still appealling to inappropriate standard regression assumptions for inference. Since the statistical package used for this last portion of the balanced effects analysis uses unweighted sample counts to calculate its degrees of freedom, the analyses so obtained are equivalent to those that would have resulted from first scaling the sampling weights so that they summed to the unweighted sample size and then using a statistical package that used the sum of the weights as its total degrees of freedom. The balanced effect F-tests along with their significance or probability levels are presented in Appendix C.

3.3 Results

3.3.1 Wald Statistic/Weight Least Squares

The design effects (DEFFs) for each domain p-value and mean score used in the Wald statistics/weighted and least squares analyses are summarized in Tables 3-3, 3-4, and 3-5. Each table presents the minimum, median, maximum and mean DEFFs for a particular NAEP item or mean score across the levels of the indicated domain defining cross-classification (i.e., Race*Sex*PARED or Sex*TOC*PARED). The design effects reported in these three tables are consistent with previous NAEP experience and tend to average around 1.4. Also, as discussed in section 2.2, the mean DEFF's given in the last column of each table are the exact quantities proposed by Rao and Scott (1981) and Fellegi (1980) for adjusting simple random sampling (SRS) based Wald Statistics chi-squareds to reflect the effects of the sample design. These are the adjustment factors used in the subsequent discussion.

As was noted in section 3.2, two different methods of analyses or hypothesis testing often used by researchers was considered within the Wald statistic/weight least squares context. The first fitted a linear model to the estimated domain statistics. Relevant hypotheses were then tested via contrasts of the estimated linear model parameters. The parameters were estimated weighting inversely proportional to the SRS covariance matrix of the domain statistics to obtain the SRS test statistics. Another set of parameter estimates was obtained by weighting by the inverse of the design based covariance matrix and the asymptotically correct test statistics were calculated. The second method of analysis evaluated hypotheses via direct contrasts of the domain statistics. Again this was first accomplished using the SRS covariance matrix to obtain the SRS test statistics, and was then repeated using the design based covariance matrix to obtain the asympotically correct tests. Results in the rest of this section will be presented separately for these two modes of analysis (i.e., contrasts of linear model coefficient and contrasts of cell means).

Table 3-3. NAEP Item Design Effects for the Race * Sex * PARED Cross-Classification

NAEP Item	Minimum DEFF	Median DEFF	Maximum DEFF	Mean DEFF
N0222A	.79	1.23	3.08	1.48
N0227A	.80	1.36	1.94	1.40
N0305C	.62	1.39	1.93	1.35
N0323A	.59	1.27	1.67	1.14
T0105A	.91	1.50	2.84	1.63
T0110A	.56	1.26	2.38	1.43
T0203A	.99	1.72	2.29	1.66
T0223A	.69	1.13	2.32	1.28
T0224A	1.00	1.31	2.82	1.47
S0108A	.63	.94	1.99	1.11
S0117A	.61	1.17	2.44	1.23
S0121A	.39	1.09	3.71	1.37
S0206A	.72	1.25	3.44	1.40
S0225A	.59	.84	1.83	.99
Average	.71	1.25	2.48	1.35

Table 3-4. NAEP Item Design Effects for the Sex * TOC * PARED Cross-Classification

NAEP Item	Minimum DEFF	Median DEFF	Maximum DEFF	Mean DEFF
NO222A	.21	1.17	2.49	1.25
NO305C	.37	1.53	2.21	1.35
TO105A	.49	1.40	4.32	1.61
TO110A	.64	1.28	3.02	1.31
TO203A	.27	1.36	4.46	1.62
TO223A	.68	1.14	2.10	1.25
SO108A	.44	1.03	2.01	1.14
SO117A	.35	1.11	2.14	1.14
SO206A	.48	1.53	4.17	1.66
SO225A	.47	.93	2.37	1.04
Average	.44	1.25	2.93	1.34

Table 3-5. Mean Scores Design Effects

Model/Age	Minimum DEFF	Median DEFF	Maximum DEFF	Mean DEFF
RACE*SEX*PARED				
9-year-olds	.57	1.45	3.32	1.50
13-year-olds	.78	1.31	2.33	1.46
17-year-olds	.49	1.09	2.57	1.16
Average	.61	1.28	2.74	1.37
SEX*TOC*PARED				
9-year-olds	.80	1.52	3.47	1.66
13-year-olds	.59	1.50	3.57	1.66
17-year-olds	.75	1.30	2.61	1.45
Average	.71	1.44	3.32	1.59

For each hypothesis test entertained in this portion of the investigation, the ratio of the SRS based Wald statistic to the asymptotically correct sample design based Wald statistic chi-squared was calculated. These ratios are another measure of the effect of the sample design and are referred to in the remaining tables as hypothesis test design effects. Two issues will be addressed by way of these test DEFFs. First, an indication of the ordinal relationship between the two test statistics will be sought. That is, does the SRS statistic tend to be generally smaller or larger than the design based chi-squared? Second, are the test DEFFs fairly constant, at least within an item or mean score? This second point is important if a simple multiplicative adjustment to the SRS test statistics is to be successful. Tables 3-6, 3-7, 3-8, present a summary of the test DEFFs for each mean or item score for the indicated cross-classification. The minimum, median, maximum and mean test design effects are shown separately for linear model coefficient contrasts (test numbers 1 through 8 in Tables 3-1 and 3-2) and cell mean contrasts (test numbers 9 through 17 in Tables 3-1 and 3-2).

The most striking feature of these three tables is the extreme instability of the test DEFFs for linear model coefficients. In virtually every case the mean is far greater than the median, indicating a skewed distribution with a long right hand tail. It appears that adjusting the SRS test statistic for the linear model coefficient contrasts will not prove fruitful because of the extreme range they cover. This may result from using the SRS covariance matrix to estimate the linear model parameters for the SRS test statistic. This process does not properly account for the correlated nature of the domain statistics and leads to less precise estimates of the model coefficients. Conversely, Tables 3-6, 3-7, and 3-8

Table 3-6. Hypothesis Test Design Effects by NAEP Item for the Race * Sex * PARED Cross-Classification

NAEP Item	Contrast of Linear Model Coefficients				Contrast of Cell Means			
	Minimum	Median	Maximum	Mean	Minimum	Median	Maximum	Mean
N0222A	.04	.82	5.42	1.41	.19	.74	1.81	.88
N0227A	.00	.57	900.26	112.96	.23	1.02	1.60	.88
N0305C	.09	.57	18.69	3.88	.62	1.33	2.40	1.38
N0323A	.00	.48	1.08	.57	.51	1.08	2.01	1.12
T0105A	.32	.99	15.73	4.02	.44	1.16	1.98	1.27
T0110A	.16	.63	1.72	.81	.56	1.18	2.18	1.19
T0203A	.10	.86	2.29	1.03	.53	1.51	2.21	1.50
T0223A	.49	5.10	284.87	45.21	.72	1.11	1.63	1.10
T0224A	.80	1.68	34.09	9.05	.65	1.10	2.41	1.27
S0108A	.03	.71	47.13	6.47	.55	.84	1.50	.93
S0117A	.19	.59	3.62	.97	.53	.75	1.75	1.00
S0121A	.00	.47	26.19	3.91	.60	.95	2.23	1.19
S0206A	.59	1.51	2.67	1.58	.59	1.10	2.09	1.12
S0225A	.34	.65	2.33	.87	.43	.92	1.09	.84
Average	.23	1.12	96.15	13.77	.51	1.06	1.92	1.12

Table 3-7. Hypothesis Test Design Effects by NAEP Item for the Sex * TOC * PARED Cross-Classification

NAEP Item	Contrast of Linear Model Coefficients				Contrast of Cell Means			
	Minimum	Median	Maximum	Mean	Minimum	Median	Maximum	Mean
N0222A	.48	4.28	55.14	11.51	.11	.48	2.82	.75
N0305C	.10	1.08	190.09	29.98	.19	.97	1.81	.89
T0105A	.04	.60	6.97	1.70	.13	.39	3.23	.98
T0110A	.37	.76	1.57	.80	.19	.55	3.41	.84
T0203A	.14	.44	3.93	.91	.27	1.08	1.84	.93
T0223A	.22	1.23	10.30	2.40	.45	.86	1.13	.77
S0108A	.02	.14	.64	.22	.10	.36	2.62	.73
S0117A	.46	.97	2.80	1.22	.03	.36	2.46	.70
S0206A	.11	.47	1.27	.54	.10	.64	1.27	.59
S0225A	.05	.75	2.98	.98	.23	.45	1.43	.60
Average	.20	1.07	27.57	5.03	.18	.61	2.20	.78

Table 3-8. Hypothesis Test Design Effects for Mean Scores

Model/Age	Contrast of Linear Model Coefficients Minimum	Median	Maximum	Mean	Contrast of Cell Means Minimum	Median	Maximum	Mean
Race*Sex*PARED								
9-year-olds	.11	.22	3.74	.85	.29	.91	1.67	1.00
13-year-olds	.09	1.86	7064.23	885.11	.59	1.19	2.23	1.26
17-year-olds	.00	.43	1.16	.56	.40	1.08	1.32	.89
Average	.07	.84	2356.38	295.51	.43	1.06	1.74	1.05
Sex*TOC*PARED								
9-year-olds	.23	.39	1.39	.55	.19	.62	2.53	.91
13-year-olds	.05	.50	1.96	.74	.17	.72	2.87	1.09
17-year-olds	.02	.65	223.55	28.54	.03	.50	1.27	.53
Average	.10	.51	75.63	9.94	.13	.61	2.22	.84

indicate that the cell mean contrast hypothesis test design effects tend to be more symetrically distributed over a narrower range than their linear model counterparts. However, they still exhibit enough variation on both sides of unity to make a simple multipicative adjustment questionable.

As indicated earlier, theoretical considerations suggest that the mean design effects presented in Tables 3-3, 3-4 and 3-5 may provide serviceable adjustments to the SRS test statistics. This conclusion is drawn into question by comparing the standard mean DEFFs in these three tables with the average test DEFFs for cell mean contrasts in Tables 3-6, 3-7, and 3-8. Almost without exception the mean test DEFFs are less than their corresponding p-value DEFF average. In addition, the mean hypothesis test DEFFs are generally near unity or less while the standard mean DEFFs are generally much greater than unity. This implies that dividing the SRS test statistic by the mean design effect will produce a test that is generally much too conservative. In fact, the adjustment suggested by Rao and Scott (1981) or Fellegi (1980) is in the wrong direction for the examples presented here.

The hypothesis test design effects are further summarized in Table 3-9 through 3-12. These four tables display the distribution of the test DEFFs over NAEP Items or mean score for each of the hypothesis tests shown in Tables 3-1 and 3-2. As was noted before, the linear model tests are very unstable. An interesting observation for the cell mean contrast test DEFFs is the distinct relationship between the number of degrees of freedom (d.f.) for the test and mean test DEFF. The larger d.f. tests have the smaller mean test DEFFs. The relationship is almost deterministic. The minimum, median and maximum test DEFFs also follow this distinct relationship. This observation is surprising in light of the eigenvalue inequality presented in section 2.2. This inequality indicates that as the number of

Table 3-9. Hypothesis Test Design Effects for NAEP Items by Test Number for the Race * Sex * PARED Cross-Classification

Test Numbers	d.f.	Minimum	Median	Maximum	Mean
Linear Model Tests					
1	4	.72	.95	2.23	1.09
2	1	.31	.97	26.19	2.88
3	1	.14	.59	900.26	86.14
4	1	.66	1.17	18.69	2.41
5	1	.01	.58	17.38	2.59
6	1	.00	.64	12.10	2.20
7	1	.00	.44	55.51	10.16
8	1	.04	.56	15.73	2.67
Average		.24	.74	131.01	13.77
Contrast Tests					
9	11	.19	.64	1.18	.72
10	1	1.07	1.62	2.41	1.71
11	1	.82	1.25	2.06	1.38
12	2	.74	1.15	2.40	1.34
13	1	.51	1.10	2.09	1.22
14	6	.44	1.10	1.94	1.07
15	6	.43	.70	1.12	.74
16	8	.37	.72	1.64	.80
17	4	.72	.95	2.23	1.09
Average		.59	1.03	1.90	1.12

81

Table 3-10. Hypothesis Test Design Effects for NAEP Items by Test Number for the Sex * TOC* PARED Cross-Classification

Test Numbers	d.f.	Minimum	Median	Maximum	Mean
Linear Model Tests					
1	6	.17	.52	1.53	.78
2	1	.09	.97	190.09	20.03
3	2	.11	.66	55.14	6.30
4	1	.42	.77	3.85	1.11
5	2	.02	.58	12.47	2.65
6	1	.04	.80	45.56	6.48
7	2	.05	.69	4.42	1.00
8	2	.03	.62	11.62	1.87
Average		.12	.70	40.59	5.03
Contrast Tests					
9	17	.03	.16	.30	.19
10	1	.35	.77	2.82	.98
11	2	.32	1.31	3.23	1.53
12	2	.57	.85	1.38	.92
13	1	.41	1.05	3.41	1.34
14	9	.13	.36	1.08	.45
15	12	.11	.24	.45	.27
16	12	.14	.54	1.13	.55
17	6	.17	.52	1.53	.78
Average		.25	.64	1.73	.78

Table 3-11. Hypothesis Test Design Effects for Mean Scores by Test Number for the Race * Sex * PARED Cross-Classification

Test Numbers	d.f.	Minimum*	Median*	Maximum*	Mean
Linear Model Tests					
1	4	.29	1.15	1.83	1.09
2	1	.41	.44	3.74	1.53
3	1	.09	.15	1.16	.47
4	1	1.01	1.06	1.90	1.32
5	1	.00	.21	3.46	1.22
6	1	.22	.42	7064.23	2354.96
7	1	.01	.11	7.91	2.68
8	1	.20	.28	1.88	.79
Average		.28	.48	885.76	295.51
Contrast Tests					
9	11	.42	.76	.81	.66
10	1	1.23	1.67	2.23	1.71
11	1	.79	1.19	1.36	1.11
12	2	1.08	1.49	1.71	1.43
13	1	.91	1.08	1.37	1.12
14	6	.97	1.13	1.32	1.14
15	6	.49	.59	.76	.61
16	8	.40	.67	.76	.61
17	4	.29	1.15	1.83	1.09
Average		.73	1.08	1.35	1.05

---

*Only three observations.

Table 3-12. Hypothesis Test Design Effects for Mean Scores by Test Number for the Sex * TOC * PARED Cross-Classification

Test Numbers	d.f.	Minimum*	Median*	Maximum*	Mean
Linear Model Tests					
1	6	.54	.64	.69	.62
2	1	.02	.05	.26	.11
3	2	.31	.61	.82	.58
4	1	.19	.47	1.39	.68
5	2	.23	.58	1.96	.92
6	1	.26	1.32	223.55	75.04
7	2	.16	.53	1.24	.64
8	2	.26	1.11	1.44	.94
Average		.25	.66	30.17	9.94
Contrast Tests					
9	17	.03	.17	.43	.21
10	1	.50	.72	1.17	.80
11	2	1.27	2.53	2.74	2.18
12	2	.66	.85	1.35	.95
13	1	.62	.86	2.87	1.45
14	9	.34	.61	.89	.61
15	12	.19	.21	.51	.30
16	12	.22	.48	.60	.43
17	6	.54	.64	.69	.62
Average		.49	.79	1.25	.84

---

*Only three observations.

contrasts simultaneously tested (i.e., degrees of freedom) increase the mean test DEFF should approach the mean design effect if the adjustment to the SRS test is effective. However, the exact opposite relationship is observed. As the d.f. increase the mean test DEFF tends to depart further from the mean DEFF. This casts further doubt on the appropriateness of the mean DEFF adjustment.

The Wald statistic/weighted least squares data was also analyzed by considering the tables in Appendix D. This appendix presents contingency tables of the number of tests which were either accepted or rejected at the five percent significance level by the sample design based test versus either the SRS test or the adjusted SRS test. Recall that the adjusted test was obtained by dividing the SRS test statistic by the appropriate mean design effect given in Tables 3-3, 3-4 or 3-5. All three test statistics were compared against the chi-squared distribution with the appropriate degrees of freedom. Appendix D was further summarized by calculating the four conditional percents of reaching an opposite conclusion for each contingency table which are reported in Tables 3-13 through 3-16. The last column of the table for cell mean contrasts of NAEP items (Table 3-13) indicates that the SRS tests are actually too conservative. This seems to be especially apparent for the Sex*TOC*PARED cross-classification. Approximately 15 percent of the Race*Sex*PARED and 32 percent of the Sex*TOC*PARED hypotheses accepted by the SRS test should have been rejected. Conversely, approximately ten percent of the hypotheses accepted by the asymptotically correct sample design based test were rejected by the SRS test. This implies that while the SRS test tends to be overly conservative, it does not follow that any hypothesis rejected by the SRS test would be rejected by the sample design based test. In addition, note that the

Table 3-13. Conditional Percent of Contrast Design Based (DB) Tests Versus Alternative Tests (AT) Reaching an Opposite Conclusion for NAEP Items

Cross-Classification	Alternate Test	Rejected by AT given accepted by DB	Accepted by DB given rejected by AT	Accepted by AT given rejected by DB	Rejected by DB given accepted by AT
Race*Sex*PARED					
	SRS				
	9-year-olds	10.0	13.3	18.8	14.3
	13-year-olds	7.1	3.4	9.7	18.8
	17-year-olds	6.3	3.6	6.9	11.8
	All ages	8.0	5.6	10.5	14.8
	Adjusted				
	9-year-olds	5.0	7.1	18.8	13.6
	13-year-olds	7.1	3.6	12.9	23.5
	17-year-olds	0.0	0.0	10.3	15.8
	All ages	4.0	2.9	13.2	17.2
Sex*TOC*PARED					
	SRS				
	9-year-olds	0.0	0.0	38.5	50.0
	13-year-olds	18.2	8.0	8.0	18.2
	17-year-olds	10.0	4.3	15.4	30.8
	All ages	11.5	5.4	17.2	32.4
	Adjusted				
	9-year-olds	0.0	0.0	61.5	61.5
	13-year-olds	0.0	0.0	28.0	38.9
	17-year-olds	10.0	4.8	23.1	40.0
	All ages	3.8	2.3	32.8	45.7

Table 3-14. Conditional Percent of Contrast Design Based (DB) Test Versus Alternative Tests (AT) Reaching an Opposite Conclusion for Mean Scores

Cross-Classification	Alternate Test	Rejected by AT given accepted by DB	Accepted by DB given rejected by AT	Accepted by AT given rejected by DB	Rejected by DB given accepted by AT
Race*Sex*PARED					
	SRS				
	9-year-olds	0.0	0.0	0.0	0.0
	13-year-olds	0.0	0.0	0.0	0.0
	17-year-olds	0.0	0.0	0.0	0.0
	All ages	0.0	0.0	0.0	0.0
	Adjusted				
	9-year-olds	0.0	0.0	14.3	33.3
	13-year-olds	0.0	0.0	0.0	0.0
	17-year-olds	0.0	0.0	0.0	0.0
	All ages	0.0	0.0	5.0	12.5
Sex*TOC*PARED					
	SRS				
	9-year-olds	0.0	0.0	14.3	33.3
	13-year-olds	33.3	14.3	0.0	0.0
	17-year-olds	0.0	0.0	0.0	0.0
	All ages	12.5	5.3	5.3	12.5
	Adjusted				
	9-year-olds	0.0	0.0	42.9	60.0
	13-year-olds	33.3	25.0	50.0	60.0
	17-year-olds	0.0	0.0	33.3	40.0
	All ages	12.5	8.3	42.1	53.3

Table 3-15. Conditional Percent of Linear Model Design Based (DB) Tests Versus Alternative Tests (AT) Reaching an Opposite Conclusion for NAEP Items

Cross-Classification	Alternate Test	Rejected by AT given accepted by DB	Accepted by DB given rejected by AT	Accepted by AT given rejected by DB	Rejected by DB given accepted by AT
Race*Sex*PARED					
	SRS				
	9-year-olds	3.8	16.7	16.7	3.8
	13-year-olds	16.7	23.5	18.8	13.0
	17-year-olds	3.4	10.0	18.2	6.7
	All ages	7.6	18.2	18.2	7.6
	Adjusted				
	9-year-olds	0.0	0.0	16.7	3.7
	13-year-olds	12.5	21.4	31.3	19.2
	17-year-olds	3.4	10.0	18.2	6.7
	All ages	5.1	13.8	24.2	9.6
Sex*TOC*PARED					
	SRS				
	9-year-olds	8.3	33.3	50.0	15.4
	13-year-olds	11.1	22.2	50.0	30.4
	17-year-olds	16.7	14.3	40.0	44.4
	All ages	11.9	19.2	44.7	31.5
	Adjusted				
	9-year-olds	0.0	0.0	75.0	20.0
	13-year-olds	0.0	0.0	57.1	30.8
	17-year-olds	16.7	15.4	45.0	47.4
	All ages	4.8	10.0	52.6	33.3

Table 3-16. Conditional Percent of Linear Model Design Based (DB) Tests Versus Alternative Tests (AT) Reaching an Opposite Conclusion for Mean Scores

Cross-Classification	Alternate Test	Rejected by AT given accepted by DB	Accepted by DB given rejected by AT	Accepted by AT given rejected by DB	Rejected by DB given accepted by AT
Race*Sex*PARED					
	SRS				
	9-year-olds	0.0	0.0	60.0	50.0
	13-year-olds	0.0	0.0	0.0	0.0
	17-year-olds	0.0	0.0	33.3	16.7
	All ages	0.0	0.0	40.0	22.2
	Adjusted				
	9-year-olds	0.0	0.0	60.0	50.0
	13-year-olds	0.0	0.0	0.0	0.0
	17-year-olds	0.0	0.0	33.3	16.7
	All ages	0.0	0.0	40.0	22.2
Sex*TOC*PARED					
	SRS				
	9-year-olds	-	0.0	50.0	100.0
	13-year-olds	0.0	0.0	20.0	25.0
	17-year-olds	20.0	33.3	33.3	20.0
	All ages	12.5	9.1	37.5	46.2
	Adjusted				
	9-year-olds	-	0.0	87.5	100.0
	13-year-olds	0.0	0.0	60.0	50.0
	17-year-olds	0.0	0.0	33.3	16.7
	All ages	0.0	0.0	68.8	57.9

conservatism observed for the SRS test is exaggerated for the adjusted test. This is a further reflection of the previous observation that the mean design effect is too large of an adjustment to divide the SRS test statistic by. The same observations can be made for the contrasts of NAEP mean scores (Table 3-14). However, the results are less dramatic. In addition Tables 3-15 and 3-16 present the results for the linear model based tests. Because of the deficiencies presented previously for this mode of analysis, these two tables are presented for completeness only.

3.3.2 Balanced Effects

As noted in section 3.1, the balanced effects analysis proceeded by fitting two different linear models to the data and then assessing the significance of each term in the model after accounting from the remaining terms. These tests are presented in Appendix C. As was done for the Wald statistic/weighted least square data, contingency tables were formed of the number of tests which were either accepted or rejected at the five percent significance level by the sample design based test versus either the sampling weighted standard regression test or the unweighted standard regression test. Again, the contingency tables were further summarized to yield Table 3-17. This table presents the four conditional percents of reaching an opposite conclusion for each contingency table. The first column of this table indicates that both of the non-sample design based testing procedures are far too liberal. These two procedures tend to reject about 20 percent too often.

Table 3-17. Conditional Percent of Balanced Effects Design Based (DB) Test Versus Alternative Tests (AT) Reaching an Opposite Conclusion

Alternate Test	Rejected by AT given accepted by DB	Accepted by DB given rejected by AT	Accepted by AT given rejected by DB	Rejected by DB given accepted by AT
Unweighted				
9-year-olds	17.6	21.4	15.4	12.5
13-year-olds	22.2	08.7	00.0	00.0
17-year-olds	08.3	05.3	00.0	00.0
All ages	15.8	10.7	03.8	05.9
Weighted				
9-year-olds	23.5	25.0	07.7	07.1
13-year-olds	22.2	08.7	00.0	00.0
17-year-olds	25.0	15.0	05.6	10.0
All ages	23.7	15.3	03.8	06.5

4. COMMENTS ON NAEP DATA AND DOCUMENTATION

Two main problems were observed with the data or documentation. First of all, the documentation contains an extensive description of the NAEP sample design and indicates that this design should be considered when analyzing the data. Unfortunately, the documentation does not indicate how this design is reflected in the data. The variable ISVARES is listed as the variance estimation code, but no indication is given as how to interpret this variable. Since this work was done at RTI, we were able to determine how this variable relates to the sample design, e.g. strata and primary sampling units. The second item that we would have found useful was a machine readable key for scoring the exercises.

REFERENCES

Anderson, T. W., and Das Gupta, S. (1963), "Some Inequalities on Characteristic Roots of Matrices," Biometrika, Volume 50.

Campbell, Cathy and Meyer, Michael (1978) "Some Properties of T Confidence Intervals for Survey Data." Proceedings of the Section on Survey Research Methods, Annual Meeting of the American Statistical Association.

Chromy, James R. (1979) "Sequential Sample Selection Methods", Proceedings of the Section on Survey Research Methods, Annual Meeting of the American Statistical Association.

Felligi, I. P. (1980), "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples," Journal of the American Statistical Association, Volume 75, Number 370.

Folsom, R. E. (1974), National Assessment Approach to Sampling Error Estimation, Draft prepared for National Assessment of Educational Progress, Project 25U-796, Research Triangle Institute.

Folsom, R. E. (1980), "U-Statistics Estimation of Variance Components for Unequal Probability Samples with Nonadditive Interviewer and Respondent Errors," Proceedings of the Section on Survey Research Methods, Annual Meeting of the American Statistical Association.

Frankel, M. R. (1971). Inference from Survey Samples. Institute for Social Research, The University of Michigan, Ann Arbor, Mi.

Fuller, Wayne A. (1975) "Regression Analysis for Sample Survey". The Indian Journal of Statistics. Volume 37. Series C.

Grizzle, J. E., Starmer, C. F., Koch, G. G. (1969), "Analyis of Categorical Data by Linear Models," Biometrics, Volume 25.

Hansen, M. H., and Hurwitz, W. N. (1943). "On the Theory of Sampling from Finite Populations" Ann. Math Stat., Volume 14.

Horvitz, D. G. and Thompson, D. J. (1952). "A Generalization of Sampling Without Replacement from a Finite Universe." Journal American Statistical Association, Volume 47.

Kish, L. (1965). Survey Sampling. John Wiley & Sons, New York.

Koch, G. G., Freeman, D. H., Freeman, J. L. (1975), "Strategies in the Multivariate Analysis of Data from Complex Surveys," International Statistical Review, Volume 43, Number 1.

Krewski, D., Rao, J. N. K. (1978), "Large Sample Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods," Proceedings of the Section on Survey Research Methods, Annual Meeting of the American Statistical Association.

Krewski D., Rao, J. N. K. (1979), "Small Sample Properties of the Linearization, Jackknife and Balanced Half-Sample Methods for Ratio Estimation in Stratified Samples," *Proceedings of the Section on Survey Research Methods*, Annual Meeting of the American Statistical Association.

Lepkowski, James M. and Landis, J. Richard (1980) "Design Effects for Linear Contrasts of Proportions and Logits". *Proceedings of the Section on Survey Research Methods*, Annual Meeting of the American Statistical Association.

Rao, J. N. K., Scott, A. J. (1979), Chi-Squared Tests for Analysis of Categorical Data from Complex Surveys," *Proceedings of the Section on Survey Research Methods*, Annual Meeting of the American Statistical Association.

Rao, J. N. K., Scott, A. J. (1981) "The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables," *Journal of the American Statistical Association*, Volume 76, Number 374.

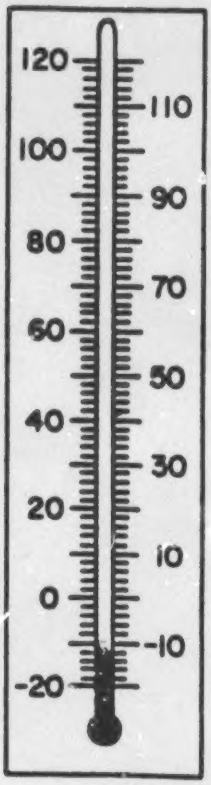
Wald, A. (1943) "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large," *Trans. Amer. Math. Soc.*, Vol. 54.

Yates, F., and Grundy, P. M. (1953) "Selection Without Replacement from within Strata with Probability Proportional to Size." *Jour. Roy. Stat. Soc.*, Volume B15.

Appendix A

NAEP Exercises

22.


120
110
100
90
80
70
60
50
40
30
20
10
0
-10
-20

What temperature is shown on this thermometer?

○ -10°

○ - 5°

○ 5°

○ 10°

○ I don't know.

Age Class 1
Package 2
Variable Name: NO222A
NAEP No.: 5-D21322
Content Area: Measurement
Unweighted Percent Correct: 79.09

STOP DO NOT CONTINUE UNTIL TOLD TO DO SO.

27. $3900 + \square = 6000$

Which one of the following is CLOSEST to the number that goes in the box?

- ○ 1000
- ○ 2000
- ○ 3000
- ○ 5000

○ I don't know.

Age Class 1
Package 2
Variable Name NO227A
NAEP No. : 5-B22745
Content Area. Variables and relationships
Unweighted Percent Correct: 23.14

STOP DO NOT CONTINUE UNTIL TOLD TO DO SO

5. A. Which is longer?

○ 2 feet

○ 1 yard

○ I don't know.

Age Class 1
Package 3
Variable Name: NO305C
NAEP No.: 5-E10003
Content Area: Other Topics
Unweighted Percent
Correct: 79.98

B. Which is heavier?

○ 17 ounces

○ 1 pound

○ I don't know.

C. Which holds more water?

○ 3 pints

○ 2 quarts

○ I don't know.

STOP DO NOT CONTINUE
UNTIL TOLD TO DO S

17.

Sarah paid $1.20 for 6 bottles of cola including the bottle deposit. If the deposit on each bottle is 5 cents what is the cost of each bottle of cola?

ANSWER ______________

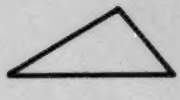
Age Class 1
Package 3
Variable Name NO317A
NAEP No. : 5-A60942
Content Area: Numbers and Numeration
Unweighted Percent Correct: 1.80

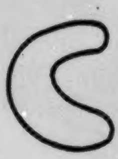
5-A60942-919-1.2.3

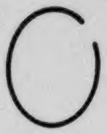
STOP DO NOT CONTINUE UNTIL TOLD TO DO SO.

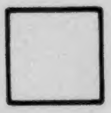
23. A. Which figure is OPEN?

Age Class 1
Package 3
Variable Name: N0323A
NAEP No. : 5-C12411
Content Area: Shape, Size and Position
Unweighted Percent
Correct: 95.37

○ 

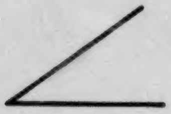
○ 

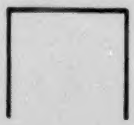
○ 

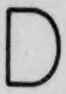
○ 

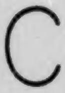
○ I don't know.

B. Which figure is CLOSED?

○ 

○ 

○ 

○ 

○ I don't know.

STOP DO NOT CONTINUE UNTIL TOLD TO DO SO.

5. $3900 + \boxed{} = 6000$

Which one of the following is CLOSEST to the number that goes in the box?

○ 1000

○ 2000

○ 3000

○ 5000

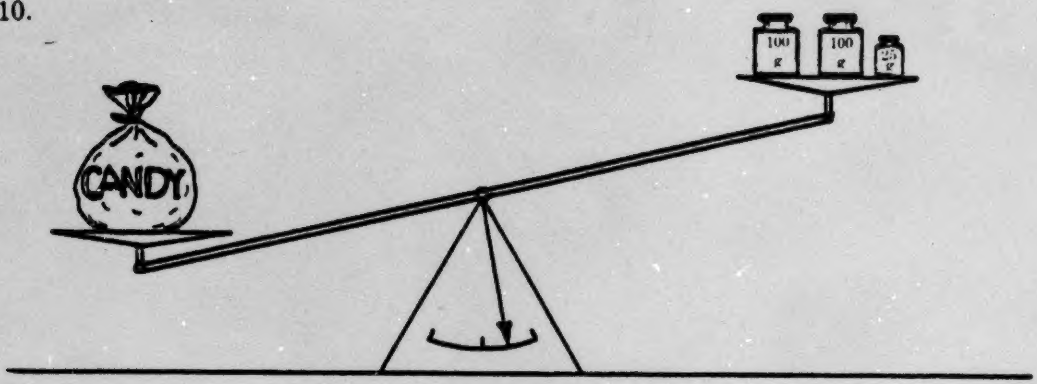
○ I don't know.

Age Class 2
Package 1
Variable Name: T0105A
NAEP No.: 5-B22745
Content Area: Variables and Relationships
Unweighted Percent Correct : 64.53

105 STOP DO NOT CONTINUE UNTIL TOLD TO DO SO.

5 B22745-91b 1 2 3

10.



What is the weight of this bag of candy?

○ 225 g

○ more than 225 g

○ less than 225 g

○ I don't know.

Age Class 2
Package 1
Variable Name: T0110A
NAEP No.: 5-D20922
Content Area: Measurement
Unweighted Percent Correct: 72.54

○○○○○○○○○○

5-D20922-919-1.2

STOP DO NOT CONTINUE UNTIL TOLD TO DO SO.

3. What does $\frac{2}{3}$ of 9 equal?

ANSWER ______________________

Age Class 2
Package 2
Variable Name: T0203A
NAPE No.: 5-C20006
Content Area: Shape, Size and Position
Unweighted Percent Correct: 47.72

5-A21341-429-1.2.3
5-C20006-1.2.3

STOP DO NOT CONTINUE UNTIL TOLD TO DO SO.

23. Kate averages 10 miles per hour on her bike. At this rate how far will she travel in 5 hours?

○ 2 miles

○ 5 miles

○ 15 miles

○ 50 miles

○ More information is needed to solve this problem.

○ I don't know.

Age Class 2
Package 2
Variable Name: TO223A
NAEP No.: 5-E20941
Content Area: Other Topics
Unweighted Percent Correct: 86.27

5-E20941-919-1.2.3

STOP DO NOT CONTINUE UNTIL TOLD TO DO SO.

24.

Sarah paid $1.20 for 6 bottles of cola including the bottle deposit. If the deposit on each bottle is 5 cents what is the cost of each bottle of cola?

ANSWER ____________

Age Class 2
Package 2
Variable Name: T0224A
NAEP No.: 5-A60942
Content Area: Numbers and Numeration
Unweighted Percent Correct: 22.24

5-A60942-919-1 2 3

STOP DO NOT CONTINUE UNTIL TOLD TO DO SO.

6. A car traveled eight kilometers in five minutes. At this speed, how many KILOMETERS could it travel in one hour?

ANSWER ______________________

Age Class 3
Package 2
Variable Name: S0206A
NAPE No: 5-C50014
Content Area: Shape, Size and Position
Unweighted Percent Correct: 54.40

STOP DO NOT CONTINUE UNTIL TOLD TO DO SO.

25.

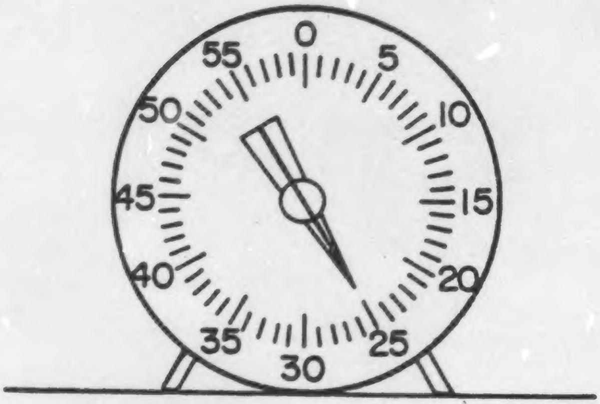
Sarah paid $1.20 for 6 bottles of cola including the bottle deposit. If the deposit on each bottle is 5 cents what is the cost of each bottle of cola?

ANSWER ____________________

Age Class 3
Package 2
Variable Name: S0225A
NAEP No.: 5-A60942
Content Area: Numbers and Numeration
Unweighted Percent Correct: 44.06

STOP DO NOT CONTINUE UNTIL TOLD TO DO SO.

8.



Betty set the timer to run for 40 minutes. It had run for 15 minutes when she discovered it should have been set to run for 50 minutes. Where should she reset the timer now to correct the mistake?

ANSWER ____________________

Age Class 3
Package 1
Variable Name: S0108A
NAEP No: 5-D94043
Content Area: Measurement
Unweighted Percent Correct: 61.29

STOP DO NOT CONTINUE UNTIL TOLD TO DO SO.

17. $3900 + \square = 6000$

Which one of the following is CLOSEST to the number that goes in the box?

○ 1000

○ 2000

○ 3000

○ 5000

○ I don't know.

Age Class 3
Package 1
Variable Name: S0117A
NAEP No.: 5-B22745
Content Area: Variables and Relationships
Unweighted Percent Correct: 85.66

STOP DO NOT CONTINUE UNTIL TOLD TO DO SO.

21. The lost dog is small and black.

A. If I see a small brown dog, then

○ it might be the lost dog.
○ it must be the lost dog.
○ it could not be the lost dog.

○ I don't know.

B. If I see a small, black dog, then

○ it might be the lost dog.
○ it must be the lost dog.
○ it could not be the lost dog.

○ I don't know.

Age Class 3
Package 1
Variable Name: S0121A
NAEP No.: 5-E50248
Content Area: Other Topics
Unweighted Percent Correct: 95.82

114

STOP DO NOT CONTINUE UNTIL TOLD TO DO SO.

5-E50248-919-1.2.3

Appendix B

Wald Statistic Chi Squareds

Table B-1. NAEP Item Wald Statistic Chi Squareds for the Race*Sex*PARED Cross-Classification

NAEP ITEM NUMBER	TEST NUMBER	D.F.	CHI SQUAREDS DESIGN BASED	CHI SQUAREDS SRS	CHI SQUAREDS ADJUSTED	SIGNIFICANCE LEVELS DESIGN BASED	SIGNIFICANCE LEVELS SRS	SIGNIFICANCE LEVELS ADJUSTED
N0222A	1	4	8.49	6.76	4.55	.0752	.1491	.3363
N0222A	2	1	0.12	0.10	0.07	.7298	.7512	.7947
N0222A	3	1	1.04	0.21	0.14	.3085	.6477	.7077
N0222A	4	1	2.20	1.89	1.28	.1377	.1687	.2587
N0222A	5	1	0.45	1.32	0.89	.5027	.2503	.3454
N0222A	6	1	2.06	0.09	0.06	.1514	.7653	.8065
N0222A	7	1	4.72	0.82	0.55	.0299	.3650	.4573
N0222A	8	1	0.32	1.76	1.18	.5691	.1849	.2767
N0222A	9	11	94.83	17.58	11.84	.0000	.0919	.3760
N0222A	10	1	0.57	0.99	0.67	.4494	.3201	.4145
N0222A	11	1	0.09	0.16	0.11	.7632	.6849	.7391
N0222A	12	2	7.12	5.28	3.55	.0285	.0714	.1691
N0222A	13	1	1.64	1.40	0.94	.2006	.2373	.3322
N0222A	14	6	9.31	6.92	4.66	.1570	.3282	.5880
N0222A	15	6	7.36	4.16	2.80	.2885	.6549	.8333
N0222A	16	8	21.77	11.41	7.68	.0054	.1797	.4651
N0222A	17	4	8.49	6.76	4.55	.0752	.1491	.3363
N0227A	1	4	4.05	3.02	2.16	.3987	.5539	.7059
N0227A	2	1	3.68	3.93	2.81	.0550	.0474	.0936
N0227A	3	1	0.00	0.73	0.52	.9773	.3933	.4704
N0227A	4	1	68.34	50.77	36.32	.0000	.0000	.0000
N0227A	5	1	1.24	0.50	0.36	.2663	.4808	.5511
N0227A	6	1	0.21	0.00	0.00	.6439	.9944	.9953
N0227A	7	1	0.35	0.05	0.04	.5535	.8158	.8438
N0227A	8	1	1.80	0.54	0.39	.1799	.4618	.5337
N0227A	9	11	598.22	138.12	98.79	.0000	.0000	.0000
N0227A	10	1	15.29	24.44	17.48	.0001	.0000	.0000
N0227A	11	1	2.11	2.45	1.75	.1466	.1178	.1859
N0227A	12	2	45.56	46.38	33.17	.0000	.0000	.0000
N0227A	13	1	0.60	0.64	0.46	.4390	.4232	.4982
N0227A	14	6	22.70	28.13	20.12	.0009	.0001	.0026
N0227A	15	6	10.15	5.05	3.61	.1185	.5379	.7294
N0227A	16	8	148.85	54.76	39.16	.0000	.0000	.0000
N0227A	17	4	4.05	3.02	2.16	.3987	.5539	.7059
N0305C	1	4	13.81	16.47	12.21	.0079	.0024	.0158
N0305C	2	1	2.13	0.66	0.49	.1445	.4149	.4826
N0305C	3	1	1.33	0.23	0.17	.2488	.6307	.6789
N0305C	4	1	0.06	1.20	0.89	.7996	.2724	.3446
N0305C	5	1	0.48	0.30	0.22	.4878	.5857	.6388
N0305C	6	1	0.31	2.91	2.16	.5791	.0880	.1419
N0305C	7	1	1.11	0.10	0.07	.2914	.7549	.7881
N0305C	8	1	0.73	0.37	0.28	.3935	.5408	.5984
N0305C	9	11	140.11	106.54	79.01	.0000	.0000	.0000
N0305C	10	1	13.49	26.62	19.74	.0002	.0000	.0000
N0305C	11	1	0.44	0.58	0.43	.5068	.4457	.5113
N0305C	12	2	2.99	7.18	5.33	.2237	.0276	.0698
N0305C	13	1	1.80	3.05	2.26	.1794	.0806	.1325
N0305C	14	6	87.93	70.44	52.24	.0000	.0000	.0000
N0305C	15	6	3.97	2.44	1.81	.6812	.8749	.9362
N0305C	16	8	17.00	27.92	20.71	.0301	.0005	.0080
N0305C	17	4	13.81	16.47	12.21	.0079	.0024	.0158

Table B-1. (continued)

NAEP ITEM NUMBER	TEST NUMBER	D.F.	CHI SQUAREDS DESIGN BASED	CHI SQUAREDS SRS	CHI SQUAREDS ADJUSTED	SIGNIFICANCE LEVELS DESIGN BASED	SIGNIFICANCE LEVELS SRS	SIGNIFICANCE LEVELS ADJUSTED
N0323A	1	4	1.76	1.90	1.66	.7805	.7549	.7980
N0323A	2	1	10.67	11.24	9.85	.0011	.0008	.0017
N0323A	3	1	0.79	0.12	0.10	.3751	.7294	.7462
N0323A	4	1	6.14	6.17	5.40	.0132	.0130	.0201
N0323A	5	1	0.29	0.08	0.07	.5919	.7824	.7961
N0323A	6	1	10.11	6.80	5.95	.0015	.0091	.0147
N0323A	7	1	0.22	0.00	0.00	.6364	.9890	.9897
N0323A	8	1	0.27	0.08	0.07	.6032	.7775	.7914
N0323A	9	11	22.46	26.48	23.19	.0211	.0055	.0165
N0323A	10	1	6.51	13.10	11.47	.0107	.0003	.0007
N0323A	11	1	0.47	0.49	0.43	.4925	.4849	.5134
N0323A	12	2	6.27	7.03	6.16	.0436	.0298	.0461
N0323A	13	1	0.00	0.00	0.00	.9639	.9743	.9759
N0323A	14	6	11.64	16.64	14.57	.0704	.0107	.0239
N0323A	15	6	5.52	5.37	4.70	.4789	.4977	.5830
N0323A	16	8	11.97	9.31	8.15	.1527	.3166	.4186
N0323A	17	4	1.76	1.90	1.66	.7805	.7549	.7980
T0105A	1	4	5.80	6.73	4.14	.2143	.1509	.3879
T0105A	2	1	10.40	8.51	5.23	.0013	.0035	.0222
T0105A	3	1	0.15	0.06	0.03	.7009	.8139	.8535
T0105A	4	1	48.51	56.09	34.48	.0000	.0000	.0000
T0105A	5	1	0.03	0.01	0.01	.8639	.9224	.9391
T0105A	6	1	0.01	0.11	0.07	.9254	.7448	.7985
T0105A	7	1	0.23	0.12	0.08	.6286	.7253	.7829
T0105A	8	1	0.01	0.11	0.07	.9319	.7347	.7905
T0105A	9	11	314.21	179.88	110.57	.0000	.0000	.0000
T0105A	10	1	24.47	45.21	27.79	.0000	.0000	.0000
T0105A	11	1	1.24	2.44	1.50	.2663	.1179	.2202
T0105A	12	2	39.22	58.07	35.70	.0000	.0000	.0000
T0105A	13	1	1.05	2.04	1.25	.3059	.1535	.2631
T0105A	14	6	114.55	49.98	30.72	.0000	.0000	.0000
T0105A	15	6	6.83	7.41	4.56	.3373	.2844	.6018
T0105A	16	8	91.33	82.79	50.89	.0000	.0000	.0000
T0105A	17	4	5.80	6.73	4.14	.2143	.1509	.3873
T0110A	1	4	4.52	6.83	4.79	.3404	.1451	.3097
T0110A	2	1	15.42	26.54	18.60	.0001	.0000	.0000
T0110A	3	1	20.95	3.32	2.33	.0000	.0682	.1269
T0110A	4	1	29.37	25.14	17.62	.0000	.0000	.0000
T0110A	5	1	14.10	4.20	2.95	.0002	.0404	.0861
T0110A	6	1	8.67	10.69	7.49	.0032	.0011	.0062
T0110A	7	1	30.87	8.06	5.65	.0000	.0045	.0175
T0110A	8	1	14.07	5.64	3.95	.0002	.0176	.0469
T0110A	9	11	97.28	96.54	67.68	.0000	.0000	.0000
T0110A	10	1	20.57	44.85	31.44	.0000	.0000	.0000
T0110A	11	1	2.77	2.64	1.85	.0959	.1042	.1737
T0110A	12	2	14.21	21.31	14.94	.0008	.0000	.0006
T0110A	13	1	0.21	0.25	0.17	.6466	.6179	.6762
T0110A	14	6	45.16	54.61	38.28	.0000	.0000	.0000
T0110A	15	6	44.39	24.78	17.37	.0000	.0034	.0080
T0110A	16	8	70.17	43.90	30.77	.0000	.0000	.0002
T0110A	17	4	4.52	6.83	4.79	.3404	.1451	.3097

Table B-1. (continued)

NAEP ITEM NUMBER	TEST NUMBER	D.F.	CHI SQUAREDS DESIGN BASED	CHI SQUAREDS SRS	CHI SQUAREDS ADJUSTED	SIGNIFICANCE LEVELS DESIGN BASED	SIGNIFICANCE LEVELS SRS	SIGNIFICANCE LEVELS ADJUSTED
T0203A	1	4	0.97	1.46	0.88	.9149	.8332	.9272
T0203A	2	1	1.54	1.36	0.82	.2150	.2432	.3646
T0203A	3	1	5.72	4.80	2.90	.0168	.0285	.0888
T0203A	4	1	13.32	30.49	18.40	.0003	.0000	.0000
T0203A	5	1	3.67	2.68	1.62	.0554	.1015	.2033
T0203A	6	1	2.42	2.60	1.57	.1199	.1069	.2164
T0203A	7	1	2.23	1.80	1.09	.1351	.1799	.2975
T0203A	8	1	0.46	0.05	0.03	.4962	.8261	.8645
T0203A	9	11	394.74	209.24	126.27	.0000	.0000	.0000
T0203A	10	1	30.87	49.39	29.80	.0000	.0000	.0000
T0203A	11	1	3.59	7.40	4.47	.0582	.0065	.0345
T0203A	12	2	13.99	30.86	18.62	.0009	.0000	.0001
T0203A	13	1	0.24	0.36	0.21	.6222	.5509	.6431
T0203A	14	6	43.51	84.27	50.85	.0000	.0000	.0000
T0203A	15	6	43.42	48.40	29.21	.0000	.0000	.0001
T0203A	16	8	50.02	54.57	32.93	.0000	.0000	.0001
T0203A	17	4	0.97	1.46	0.88	.9149	.8332	.9270
T0223A	1	4	11.92	8.63	6.72	.0180	.0711	.1513
T0223A	2	1	2.13	1.05	0.82	.1440	.3054	.3656
T0223A	3	1	0.02	4.89	3.81	.8958	.0271	.0510
T0223A	4	1	4.03	5.46	4.26	.0446	.0194	.0391
T0223A	5	1	0.59	5.20	4.05	.4435	.0226	.0441
T0223A	6	1	1.49	0.91	0.71	.2226	.3399	.3996
T0223A	7	1	0.10	5.60	4.36	.7507	.0179	.0367
T0223A	8	1	0.57	5.32	4.14	.4495	.0211	.0418
T0223A	9	11	61.54	69.13	53.86	.0000	.0000	.0000
T0223A	10	1	14.63	23.85	18.58	.0001	.0000	.0000
T0223A	11	1	0.28	0.27	0.21	.5967	.6007	.6441
T0223A	12	2	8.18	12.11	9.43	.0167	.0023	.0089
T0223A	13	1	3.98	3.61	2.81	.0461	.0574	.0935
T0223A	14	6	37.66	41.91	32.65	.0000	.0000	.0000
T0223A	15	6	9.17	6.77	5.27	.1643	.3429	.5093
T0223A	16	8	24.34	29.27	22.80	.0020	.0003	.0036
T0223A	17	4	11.92	8.63	6.72	.0180	.0711	.1513
T0224A	1	4	10.67	8.51	5.77	.0306	.0746	.2167
T0224A	2	1	0.09	0.19	0.13	.7648	.6616	.7184
T0224A	3	1	0.03	0.40	0.27	.8694	.5272	.6024
T0224A	4	1	18.84	22.91	15.55	.0000	.0000	.0001
T0224A	5	1	0.10	1.65	1.12	.7576	.1983	.2893
T0224A	6	1	9.85	9.20	6.24	.0017	.0024	.0125
T0224A	7	1	0.05	1.70	1.15	.8234	.1926	.2831
T0224A	8	1	1.93	2.04	1.38	.1648	.1532	.2393
T0224A	9	11	245.65	201.04	136.43	.0000	.0000	.0000
T0224A	10	1	29.35	70.86	48.09	.0000	.0000	.0000
T0224A	11	1	0.13	0.27	0.18	.7142	.6050	.6701
T0224A	12	2	21.37	23.54	15.97	.0000	.0000	.0003
T0224A	13	1	0.60	0.68	0.46	.4395	.4104	.4977
T0224A	14	6	75.95	115.41	78.32	.0000	.0000	.0000
T0224A	15	6	15.22	15.44	10.48	.0186	.0171	.1058
T0224A	16	8	76.96	50.40	34.20	.0000	.0000	.0000
T0224A	17	4	10.67	8.51	5.77	.0306	.0746	.2167

Table B-1. (continued)

NAEP ITEM NUMBER	TEST NUMBER	D.F.	CHI SQUAREDS			SIGNIFICANCE LEVELS		
			DESIGN BASED	SRS	ADJUSTED	DESIGN BASED	SRS	ADJUSTED
S0108A	1	4	7.65	5.67	5.12	.1051	.2255	.2753
S0108A	2	1	30.47	14.63	13.21	.0000	.0001	.0003
S0108A	3	1	1.14	0.16	0.15	.2852	.6884	.7031
S0108A	4	1	27.46	20.53	18.54	.0000	.0000	.0000
S0108A	5	1	0.03	0.05	0.05	.8688	.8227	.8314
S0108A	6	1	2.25	0.06	0.06	.1337	.7991	.8089
S0108A	7	1	0.00	0.19	0.17	.9494	.6630	.6787
S0108A	8	1	0.45	0.30	0.27	.5036	.5829	.6017
S0108A	9	11	154.91	180.71	163.23	.0000	.0000	.0000
S0108A	10	1	51.82	77.47	69.98	.0000	.0000	.0000
S0108A	11	1	5.15	4.23	3.82	.0233	.0397	.0506
S0108A	12	2	27.38	23.28	21.03	.0000	.0000	.0000
S0108A	13	1	0.22	0.12	0.11	.6363	.7255	.7386
S0108A	14	6	73.31	87.82	79.32	.0000	.0000	.0000
S0108A	15	6	9.33	7.87	7.11	.1557	.2480	.3111
S0108A	16	8	46.39	34.49	31.15	.0000	.0000	.0001
S0108A	17	4	7.65	5.67	5.12	.1051	.2255	.2753
S0117A	1	4	9.55	7.08	5.76	.0486	.1315	.2180
S0117A	2	1	5.03	5.97	4.85	.0249	.0146	.0277
S0117A	3	1	2.74	1.41	1.14	.0977	.2353	.2847
S0117A	4	1	9.13	6.05	4.92	.0025	.0139	.0266
S0117A	5	1	3.34	0.64	0.52	.0676	.4234	.4705
S0117A	6	1	0.01	0.05	0.04	.9083	.8264	.8433
S0117A	7	1	2.56	1.34	1.09	.1096	.2471	.2968
S0117A	8	1	2.81	0.85	0.69	.0935	.3577	.4071
S0117A	9	11	136.21	91.62	74.45	.0000	.0000	.0000
S0117A	10	1	26.75	42.60	34.62	.0000	.0000	.0000
S0117A	11	1	0.03	0.04	0.03	.8709	.8402	.8558
S0117A	12	2	10.42	9.32	7.57	.0055	.0095	.0227
S0117A	13	1	0.01	0.02	0.02	.9134	.8856	.8969
S0117A	14	6	65.07	48.65	39.53	.0000	.0000	.0000
S0117A	15	6	7.04	3.73	3.03	.3169	.7136	.8052
S0117A	16	8	44.33	23.55	19.13	.0000	.0027	.0142
S0117A	17	4	9.55	7.08	5.76	.0486	.1315	.2180
S0121A	1	4	1.01	2.25	1.65	.9081	.6895	.8002
S0121A	2	1	0.01	0.33	0.24	.9103	.5643	.6219
S0121A	3	1	1.21	0.80	0.59	.2722	.3700	.4432
S0121A	4	1	1.47	2.50	1.83	.2247	.1136	.1759
S0121A	5	1	0.84	0.00	0.00	.3598	.9485	.9560
S0121A	6	1	0.16	0.02	0.02	.6867	.8838	.9005
S0121A	7	1	0.08	0.02	0.02	.7837	.8840	.9007
S0121A	8	1	3.74	0.15	0.11	.0530	.6979	.7399
S0121A	9	11	21.48	12.98	9.50	.0287	.2945	.5760
S0121A	10	1	1.92	2.28	1.67	.1655	.1310	.1965
S0121A	11	1	2.72	4.54	3.32	.0994	.0330	.0682
S0121A	12	2	1.70	3.05	2.23	.4275	.2176	.3276
S0121A	13	1	0.59	0.51	0.38	.4406	.4736	.5399
S0121A	14	6	6.45	3.88	2.84	.3749	.6934	.8292
S0121A	15	6	9.45	7.44	5.44	.1496	.2821	.4884
S0121A	16	8	6.59	6.24	4.56	.5818	.6205	.8030
S0121A	17	4	1.01	2.25	1.65	.9081	.6895	.8002

Table B-1. (continued)

NAEP ITEM NUMBER	TEST NUMBER	D.F.	CHI SQUAREDS DESIGN BASED	CHI SQUAREDS SRS	CHI SQUAREDS ADJUSTED	SIGNIFICANCE LEVELS DESIGN BASED	SIGNIFICANCE LEVELS SRS	SIGNIFICANCE LEVELS ADJUSTED
S0206A	1	4	0.89	1.07	0.77	.9257	.8987	.9431
S0206A	2	1	14.43	11.52	8.22	.0001	.0007	.0041
S0206A	3	1	5.66	11.61	8.28	.0173	.0007	.0040
S0206A	4	1	40.44	48.10	34.33	.0000	.0000	.0000
S0206A	5	1	0.25	0.45	0.32	.6203	.5031	.5716
S0206A	6	1	1.65	0.98	0.70	.1992	.3234	.4041
S0206A	7	1	1.67	3.82	2.73	.1964	.0505	.0985
S0206A	8	1	0.12	0.33	0.23	.7269	.5681	.6297
S0206A	9	11	683.82	406.62	290.21	.0000	.0000	.0000
S0206A	10	1	73.81	116.97	83.48	.0000	.0000	.0000
S0206A	11	1	13.64	15.02	10.72	.0002	.0001	.0011
S0206A	12	2	40.32	47.25	33.72	.0000	.0000	.0000
S0206A	13	1	0.08	0.16	0.11	.7825	.6901	.7362
S0206A	14	6	131.82	127.38	90.91	.0000	.0000	.0000
S0206A	15	6	36.50	23.76	16.96	.0000	.0006	.0094
S0206A	16	8	122.41	84.94	60.62	.0000	.0000	.0000
S0206A	17	4	0.89	1.07	0.77	.9257	.8987	.9431
S0225A	1	4	12.29	10.06	10.16	.0153	.0395	.0378
S0225A	2	1	1.70	3.96	4.01	.1919	.0465	.0454
S0225A	3	1	0.75	0.52	0.53	.3859	.4698	.4675
S0225A	4	1	20.49	26.47	26.74	.0000	.0000	.0000
S0225A	5	1	3.80	2.05	2.07	.0513	.1522	.1501
S0225A	6	1	8.33	2.80	2.83	.0039	.0943	.0925
S0225A	7	1	1.73	0.62	0.63	.1883	.4300	.4277
S0225A	8	1	1.21	0.73	0.74	.2714	.3921	.3896
S0225A	9	11	516.97	311.19	314.43	.0000	.0000	.0000
S0225A	10	1	85.29	91.22	92.17	.0000	.0000	.0000
S0225A	11	1	18.13	16.67	16.84	.0000	.0000	.0000
S0225A	12	2	25.68	26.72	27.00	.0000	.0000	.0000
S0225A	13	1	6.65	7.08	7.15	.0099	.0078	.0075
S0225A	14	6	104.31	113.36	114.55	.0000	.0000	.0000
S0225A	15	6	62.55	26.63	26.91	.0000	.0002	.0002
S0225A	16	8	111.85	57.14	57.73	.0000	.0000	.0000
S0225A	17	4	12.29	10.06	10.16	.0153	.0395	.0378

Table B-2. NAEP Item Wald Statistic Chi Squareds for the Sex*TOC*PARED Cross-Classification

NAEP ITEM NUMBER	TEST NUMBER	D.F.	CHI SQUAREDS DESIGN BASED	CHI SQUAREDS SRS	CHI SQUAREDS ADJUSTED	SIGNIFICANCE LEVELS DESIGN BASED	SIGNIFICANCE LEVELS SRS	SIGNIFICANCE LEVELS ADJUSTED
N0222A	1	6	15.78	7.50	5.98	.0150	.2769	.4256
N0222A	2	1	0.27	0.81	0.65	.6022	.3674	.4210
N0222A	3	2	0.03	1.46	1.16	.9869	.4826	.5595
N0222A	4	1	4.76	4.13	3.29	.0291	.0422	.0697
N0222A	5	2	0.33	4.14	3.30	.8471	.1263	.1922
N0222A	6	1	0.51	2.13	1.69	.4732	.1448	.1931
N0222A	7	2	0.18	0.80	0.64	.9134	.6700	.7268
N0222A	8	2	0.34	3.99	3.18	.8421	.1358	.2037
N0222A	9	17	272.17	30.69	24.46	.0000	.0218	.1075
N0222A	10	1	0.01	0.02	0.01	.9396	.8991	.9099
N0222A	11	2	2.93	2.76	2.20	.2310	.2513	.3326
N0222A	12	2	15.62	10.29	8.20	.0004	.0058	.0165
N0222A	13	1	10.92	5.25	4.18	.0010	.0220	.0409
N0222A	14	9	37.64	13.70	10.92	.0000	.1334	.2814
N0222A	15	12	87.61	17.06	13.60	.0000	.1472	.3271
N0222A	16	12	24.49	17.69	14.09	.0180	.1255	.2947
N0222A	17	6	15.78	7.50	5.98	.0150	.2769	.4256
N0305C	1	6	14.50	22.05	16.28	.0245	.0012	.0123
N0305C	2	1	0.61	1.16	0.85	.9378	.2820	.3553
N0305C	3	2	5.29	0.55	0.41	.0710	.7579	.8149
N0305C	4	1	3.39	5.06	3.73	.0656	.0245	.0533
N0305C	5	2	15.28	1.73	1.27	.0005	.4218	.5288
N0305C	6	1	0.03	1.18	0.87	.8722	.2775	.3508
N0305C	7	2	3.69	1.02	0.76	.1581	.5995	.6854
N0305C	8	2	2.01	1.35	1.00	.3663	.5089	.6073
N0305C	9	17	248.08	47.01	34.70	.0000	.0001	.0068
N0305C	10	1	0.49	0.59	0.44	.4847	.4422	.5091
N0305C	11	2	0.44	0.81	0.59	.8010	.6686	.7429
N0305C	12	2	7.42	7.19	5.31	.0245	.0275	.0704
N0305C	13	1	0.11	0.13	0.10	.7366	.7167	.7553
N0305C	14	9	20.74	6.70	4.95	.0139	.6681	.8389
N0305C	15	12	100.11	23.08	17.04	.0000	.0270	.1481
N0305C	16	12	66.27	38.15	28.16	.0000	.0001	.0052
N0305C	17	6	14.50	22.05	16.28	.0245	.0012	.0123
T0105A	1	6	190.59	33.11	20.55	.0000	.0000	.0022
T0105A	2	1	15.54	1.44	0.89	.0001	.2301	.3444
T0105A	3	2	5.83	5.92	3.68	.0543	.0517	.1590
T0105A	4	1	14.51	55.83	34.66	.0001	.0000	.0000
T0105A	5	2	0.85	5.91	3.67	.6546	.0521	.1597
T0105A	6	1	14.27	0.52	0.32	.0002	.4712	.5703
T0105A	7	2	10.69	1.48	0.92	.0048	.4779	.6324
T0105A	8	2	5.02	6.63	4.11	.0812	.0363	.1278
T0105A	9	17	1103.92	156.36	97.06	.0000	.0000	.0000
T0105A	10	1	2.93	4.24	2.63	.0867	.0394	.1047
T0105A	11	2	2.84	9.19	5.70	.2413	.0101	.0578
T0105A	12	2	49.74	67.02	41.60	.0000	.0000	.0000
T0105A	13	1	0.06	0.10	0.06	.8056	.7547	.8056
T0105A	14	9	234.58	30.39	18.87	.0000	.0004	.0264
T0105A	15	12	117.33	45.94	28.52	.0000	.0000	.0045
T0105A	16	12	354.06	124.54	77.30	.0000	.0000	.0000
T0105A	17	6	190.59	33.11	20.55	.0000	.0000	.0022

Table B-2. (continued)

NAEP ITEM NUMBER	TEST NUMBER	D.F.	CHI SQUAREDS DESIGN BASED	CHI SQUAREDS SRS	CHI SQUAREDS ADJUSTED	SIGNIFICANCE LEVELS DESIGN BASED	SIGNIFICANCE LEVELS SRS	SIGNIFICANCE LEVELS ADJUSTED
T0110A	1	6	4.19	2.09	1.60	.6514	.9112	.9529
T0110A	2	1	0.13	0.21	0.16	.7135	.6457	.6879
T0110A	3	2	16.68	16.14	12.32	.0002	.0003	.0021
T0110A	4	1	25.66	17.50	13.36	.0000	.0000	.0003
T0110A	5	2	7.36	2.73	2.09	.0252	.2548	.3521
T0110A	6	1	2.47	2.58	1.97	.1164	.1081	.1604
T0110A	7	2	5.57	4.68	3.57	.0617	.0962	.1674
T0110A	8	2	7.57	3.35	2.56	.0227	.1868	.2779
T0110A	9	17	445.48	82.67	63.11	.0000	.0000	.0000
T0110A	10	1	14.21	5.14	3.93	.0002	.0233	.0475
T0110A	11	2	31.12	31.90	24.35	.0000	.0000	.0000
T0110A	12	2	27.58	18.15	13.85	.0000	.0001	.0010
T0110A	13	1	0.03	0.11	0.08	.8595	.7439	.7753
T0110A	14	9	35.32	19.42	14.82	.0001	.0219	.0959
T0110A	15	12	183.60	41.33	31.55	.0000	.0000	.0016
T0110A	16	12	40.20	26.90	20.54	.0001	.0080	.0576
T0110A	17	6	4.19	2.09	1.60	.6514	.9112	.9529
T0203A	1	6	11.78	5.73	3.53	.0671	.4537	.7396
T0203A	2	1	2.69	1.03	0.64	.1012	.3092	.4248
T0203A	3	2	12.96	5.19	3.20	.0015	.0746	.2021
T0203A	4	1	56.15	51.33	31.62	.0000	.0000	.0000
T0203A	5	2	0.35	1.37	0.85	.8397	.5033	.6551
T0203A	6	1	0.43	0.06	0.04	.5143	.8088	.8493
T0203A	7	2	7.28	5.69	3.50	.0262	.0582	.1735
T0203A	8	2	0.55	0.12	0.07	.7603	.9417	.9637
T0203A	9	17	525.61	142.34	87.70	.0000	.0000	.0000
T0203A	10	1	5.44	6.00	3.70	.0197	.0143	.0545
T0203A	11	2	0.39	0.73	0.45	.8210	.6957	.7997
T0203A	12	2	39.69	50.15	30.90	.0000	.0000	.0000
T0203A	13	1	0.23	0.31	0.19	.6339	.5754	.6602
T0203A	14	9	38.83	41.92	25.83	.0000	.0000	.0022
T0203A	15	12	55.45	18.35	11.30	.0000	.1055	.5030
T0203A	16	12	147.27	91.34	56.28	.0000	.0000	.0000
T0203A	17	6	11.78	5.73	3.53	.0671	.4537	.7396
T0223A	1	6	24.96	13.69	10.92	.0003	.0333	.0908
T0223A	2	1	1.86	2.50	2.00	.1724	.1136	.1576
T0223A	3	2	0.96	3.61	2.88	.6179	.1641	.2365
T0223A	4	1	5.48	6.52	5.20	.0192	.0107	.0225
T0223A	5	2	2.05	0.46	0.36	.3597	.7961	.8337
T0223A	6	1	0.41	4.27	3.41	.5198	.0388	.0650
T0223A	7	2	2.67	3.37	2.69	.2634	.1856	.2609
T0223A	8	2	3.59	2.04	1.63	.1661	.3601	.4428
T0223A	9	17	106.90	53.68	42.82	.0000	.0000	.0005
T0223A	10	1	0.05	0.04	0.03	.8244	.8344	.8519
T0223A	11	2	1.53	1.39	1.11	.4652	.4990	.5744
T0223A	12	2	12.49	10.71	8.54	.0020	.0047	.0140
T0223A	13	1	0.83	0.77	0.62	.3611	.3788	.4318
T0223A	14	9	22.55	16.21	12.93	.0073	.0626	.1657
T0223A	15	12	47.94	21.63	17.25	.0000	.0419	.1403
T0223A	16	12	37.83	42.60	33.98	.0002	.0000	.0007
T0223A	17	6	24.96	13.69	10.92	.0003	.0333	.0908

Table B-2. (continued)

NAEP ITEM NUMBER	TEST NUMBER	D.F.	CHI SQUAREDS DESIGN BASED	CHI SQUAREDS SRS	CHI SQUAREDS ADJUSTED	SIGNIFICANCE LEVELS DESIGN BASED	SIGNIFICANCE LEVELS SRS	SIGNIFICANCE LEVELS ADJUSTED
S0108A	1	6	41.38	13.94	12.27	.0000	.0303	.0562
S0108A	2	1	4.97	0.51	0.45	.0258	.4762	.5039
S0108A	3	2	54.99	9.75	8.58	.0000	.0076	.0137
S0108A	4	1	121.52	77.37	68.09	.0000	.0000	.0000
S0108A	5	2	8.41	0.17	0.15	.0149	.9194	.9287
S0108A	6	1	2.32	0.19	0.17	.1274	.6601	.6799
S0108A	7	2	30.22	11.94	10.51	.0000	.0026	.0052
S0108A	8	2	4.91	0.16	0.14	.0859	.9226	.9315
S0108A	9	17	1085.28	112.83	99.39	.0000	.0000	.0000
S0108A	10	1	3.00	1.08	0.95	.0832	.2993	.3302
S0108A	11	2	2.69	7.04	6.20	.2609	.0296	.0452
S0108A	12	2	56.34	77.56	68.26	.0000	.0000	.0000
S0108A	13	1	2.24	1.43	1.26	.1344	.2317	.2619
S0108A	14	9	23.19	8.14	7.16	.0058	.5204	.6203
S0108A	15	12	111.57	33.51	29.49	.0000	.0008	.0033
S0108A	16	12	193.05	98.54	86.73	.0000	.0000	.0000
S0108A	17	6	41.38	13.94	12.27	.0000	.0303	.0562
S0117A	1	6	18.47	28.19	24.72	.0052	.0001	.0004
S0117A	2	1	10.77	6.31	5.53	.0010	.0120	.0187
S0117A	3	2	8.25	3.83	3.36	.0161	.1470	.1862
S0117A	4	1	39.17	22.74	19.94	.0000	.0000	.0000
S0117A	5	2	16.51	11.44	10.03	.0003	.0033	.0066
S0117A	6	1	2.02	5.66	4.96	.1551	.0174	.0260
S0117A	7	2	2.03	2.52	2.21	.3619	.2838	.3315
S0117A	8	2	5.42	10.18	8.92	.0665	.0062	.0115
S0117A	9	17	2562.63	72.84	63.87	.0000	.0000	.0000
S0117A	10	1	7.21	4.41	3.86	.0073	.0358	.0493
S0117A	11	2	9.95	3.17	2.78	.0069	.2051	.2493
S0117A	12	2	48.63	27.68	24.27	.0000	.0000	.0000
S0117A	13	1	0.02	0.05	0.05	.8842	.8194	.8307
S0117A	14	9	48.46	12.29	10.78	.0000	.1974	.2914
S0117A	15	12	201.02	35.78	31.37	.0000	.0004	.0017
S0117A	16	12	179.17	64.90	56.90	.0000	.0000	.0000
S0117A	17	6	18.47	28.19	24.72	.0052	.0001	.0004
S0206A	1	6	7.72	9.81	5.89	.2590	.1330	.4356
S0206A	2	1	27.44	3.71	2.23	.0000	.0542	.1357
S0206A	3	2	2.29	1.94	1.16	.3184	.3797	.5590
S0206A	4	1	214.11	88.87	53.38	.0000	.0000	.0000
S0206A	5	2	32.41	15.24	9.15	.0000	.0005	.0103
S0206A	6	1	16.90	1.85	1.11	.0000	.1733	.2913
S0206A	7	2	4.05	2.45	1.47	.1320	.2931	.4785
S0206A	8	2	27.79	13.26	7.96	.0000	.0013	.0187
S0206A	9	17	2696.03	293.69	176.40	.0000	.0000	.0000
S0206A	10	1	8.36	5.38	3.23	.0038	.0203	.0721
S0206A	11	2	0.30	0.35	0.21	.8612	.8385	.8995
S0206A	12	2	127.86	83.99	50.45	.0000	.0000	.0000
S0206A	13	1	0.38	0.34	0.20	.5384	.5592	.6508
S0206A	14	9	178.38	51.24	30.78	.0000	.0008	.0003
S0206A	15	12	361.28	37.89	22.76	.0000	.0002	.0298
S0206A	16	12	1763.83	240.56	144.49	.0000	.0000	.0000
S0206A	17	6	7.72	9.81	5.89	.2590	.1330	.4356

Table B-2. (continued)

NAEP ITEM NUMBER	TEST NUMBER	D.F.	CHI SQUAREDS DESIGN BASED	CHI SQUAREDS SRS	CHI SQUAREDS ADJUSTED	SIGNIFICANCE LEVELS DESIGN BASED	SIGNIFICANCE LEVELS SRS	SIGNIFICANCE LEVELS ADJUSTED
S 225A	1	6	7.70	7.20	6.90	.2611	.3031	.3305
S 225A	2	1	0.20	0.60	0.57	.6544	.4397	.4493
S 225A	3	2	16.09	1.84	1.77	.0003	.3979	.4134
S 225A	4	1	95.10	48.18	46.18	.0000	.0000	.0000
S 225A	5	2	1.22	1.52	1.46	.5435	.4672	.4823
S 225A	6	1	6.75	3.78	3.62	.0094	.0518	.0570
S 225A	7	2	20.62	1.06	1.02	.0000	.5883	.6014
S 225A	8	2	1.09	1.55	1.49	.5790	.4606	.4757
S 225A	9	17	592.72	133.88	128.31	.0000	.0000	.0000
S 225A	10	1	19.46	6.86	6.57	.0000	.0088	.0104
S 225A	11	2	1.78	2.55	2.45	.4098	.2789	.2941
S 225A	12	2	57.13	47.96	45.97	.0000	.0000	.0000
S 225A	13	1	1.61	0.66	0.63	.2049	.4161	.4260
S 225A	14	9	59.05	27.55	26.41	.0000	.0011	.0018
S 225A	15	12	24.94	6.26	6.03	.0151	.9022	.9166
S 225A	16	12	233.74	105.97	101.56	.0000	.0000	.0000
S 225A	17	6	7.70	7.20	6.90	.2611	.3031	.3305

Table B-3. Mean Score Wald Statistic Chi Squareds for the Race*Sex* PARED Cross-Classification

			CHI SQUAREDS			SIGNIFICANCE LEVELS		
NAEP ITEM NUMBER	TEST NUMBER	D.F.	DESIGN BASED	SRS	ADJUSTED	DESIGN BASED	SRS	ADJUSTED
AGE 9	1	4	54.14	15.88	10.59	.0000	.0032	.0315
AGE 9	2	1	0.93	3.47	2.31	.3353	.0626	.1282
AGE 9	3	1	7.17	1.09	0.72	.0074	.2976	.3949
AGE 9	4	1	18.63	35.40	23.62	.0000	.0000	.0000
AGE 9	5	1	2.59	0.55	0.36	.1073	.4597	.5459
AGE 9	6	1	4.59	1.03	0.69	.0321	.3102	.4072
AGE 9	7	1	5.88	0.63	0.42	.0153	.4275	.5169
AGE 9	8	1	2.79	0.57	0.38	.0946	.4498	.5371
AGE 9	9	11	204.52	165.65	110.52	.0000	.0000	.0000
AGE 9	10	1	28.64	47.86	31.93	.0000	.0000	.0000
AGE 9	11	1	1.10	1.50	1.00	.2938	.2209	.3174
AGE 9	12	2	32.26	48.17	32.14	.0000	.0000	.0000
AGE 9	13	1	5.22	4.74	3.16	.0223	.0295	.0755
AGE 9	14	6	76.10	74.36	49.61	.0000	.0000	.0000
AGE 9	15	6	5.09	3.86	2.57	.5324	.6962	.8603
AGE 9	16	8	83.20	63.10	42.10	.0000	.0000	.0000
AGE 9	17	4	54.14	15.88	10.59	.0000	.0032	.0315
AGE13	1	4	1.74	3.19	2.18	.7827	.5273	.7021
AGE13	2	1	59.33	24.52	16.81	.0000	.0000	.0000
AGE13	3	1	0.16	0.02	0.01	.6899	.9023	.9191
AGE13	4	1	97.02	103.15	70.70	.0000	.0000	.0000
AGE13	5	1	0.17	0.58	0.40	.6810	.4447	.5269
AGE13	6	1	0.00	0.01	0.01	.9990	.9184	.9324
AGE13	7	1	0.03	0.21	0.15	.8694	.6437	.7018
AGE13	8	1	0.02	0.03	0.02	.8945	.8556	.8803
AGE13	9	11	609.23	462.33	316.88	.0000	.0000	.0000
AGE13	10	1	75.40	168.38	115.41	.0000	.0000	.0000
AGE13	11	1	1.24	1.48	1.01	.2651	.2239	.3140
AGE13	12	2	61.34	104.64	71.72	.0000	.0000	.0000
AGE13	13	1	0.20	0.28	0.19	.6539	.5992	.6635
AGE13	14	6	159.30	180.27	123.56	.0000	.0000	.0000
AGE13	15	6	31.93	18.71	12.82	.0000	.0047	.0459
AGE13	16	8	226.51	151.74	104.00	.0000	.0000	.0000
AGE13	17	4	1.74	3.19	2.18	.7827	.5273	.7021
AGE17	1	4	2.80	3.22	2.77	.5924	.5215	.5973
AGE17	2	1	57.01	24.96	21.45	.0000	.0000	.0000
AGE17	3	1	2.44	2.83	2.44	.1181	.0923	.1186
AGE17	4	1	81.54	82.43	70.84	.0000	.0000	.0000
AGE17	5	1	0.09	0.00	0.00	.7661	.9841	.9852
AGE17	6	1	4.68	1.96	1.69	.0306	.1610	.1938
AGE17	7	1	0.06	0.00	0.00	.8144	.9829	.9841
AGE17	8	1	0.49	0.14	0.12	.4840	.7096	.7299
AGE17	9	11	1735.18	732.85	629.82	.0000	.0000	.0000
AGE17	10	1	211.20	261.44	224.68	.0000	.0000	.0000
AGE17	11	1	24.29	19.16	16.47	.0000	.0000	.0000
AGE17	12	2	74.93	80.89	69.52	.0000	.0000	.0000
AGE17	13	1	2.27	2.44	2.10	.1319	.1180	.1473
AGE17	14	6	223.27	294.93	253.47	.0000	.0000	.0000
AGE17	15	6	55.23	26.99	23.20	.0000	.0001	.0007
AGE17	16	8	356.80	142.30	122.30	.0000	.0000	.0000
AGE17	17	4	2.80	3.22	2.77	.5924	.5215	.5973

Table B-4. Mean Score Wald Statistic Chi Squareds for the Sex*TOC*PARED Cross-Classification

			CHI SQUAREDS			SIGNIFICANCE LEVELS		
NAEP ITEM NUMBER	TEST NUMBER	D.F.	DESIGN BASED	SRS	ADJUSTED	DESIGN BASED	SRS	ADJUSTED
AGE09	1	6	22.53	14.36	8.67	.0010	.0259	.1930
AGE09	2	1	7.75	2.04	1.23	.0054	.1531	.2670
AGE09	3	2	9.16	7.55	4.56	.0102	.0230	.1024
AGE09	4	1	37.40	52.09	31.45	.0000	.0000	.0000
AGE09	5	2	11.46	2.65	1.60	.0032	.2658	.4493
AGE09	6	1	13.51	3.49	2.11	.0002	.0619	.1468
AGE09	7	2	12.04	6.33	3.82	.0024	.0422	.1480
AGE09	8	2	22.48	5.93	3.58	.0000	.0515	.1668
AGE09	9	17	274.45	118.98	71.84	.0000	.0000	.0000
AGE09	10	1	0.01	0.01	0.01	.9308	.9252	.9418
AGE09	11	2	1.57	3.96	2.39	.4569	.1379	.3023
AGE09	12	2	42.96	57.90	34.96	.0000	.0000	.0000
AGE09	13	1	10.60	6.57	3.96	.0011	.0104	.0465
AGE09	14	9	22.44	13.76	8.31	.0076	.1310	.5032
AGE09	15	12	128.46	24.93	15.05	.0000	.0152	.2386
AGE09	16	12	173.84	104.41	63.04	.0000	.0000	.0000
AGE09	17	6	22.53	14.36	8.67	.0010	.0259	.1930
AGE13	1	6	25.92	13.91	8.38	.0002	.0306	.2116
AGE13	2	1	3.87	0.19	0.12	.0491	.6609	.7335
AGE13	3	2	39.25	12.29	7.40	.0000	.0021	.0247
AGE13	4	1	299.09	139.61	84.09	.0000	.0000	.0000
AGE13	5	2	1.54	3.01	1.81	.4633	.2219	.4038
AGE13	6	1	1.21	1.59	0.96	.2720	.2067	.3271
AGE13	7	2	39.49	6.28	3.78	.0000	.0433	.1510
AGE13	8	2	3.12	3.46	2.08	.2104	.1774	.3528
AGE13	9	17	1586.04	274.84	165.55	.0000	.0000	.0000
AGE13	10	1	2.39	1.72	1.03	.1222	.1903	.3094
AGE13	11	2	5.88	16.12	9.71	.0528	.0003	.0078
AGE13	12	2	159.26	135.15	81.41	.0000	.0000	.0000
AGE13	13	1	0.01	0.03	0.02	.9135	.8540	.8864
AGE13	14	9	29.59	26.23	15.80	.0005	.0019	.0712
AGE13	15	12	66.24	34.03	20.50	.0000	.0007	.0583
AGE13	16	12	451.30	215.98	130.10	.0000	.0000	.0000
AGE13	17	6	25.92	13.91	8.38	.0002	.0306	.2116
AGE17	1	6	9.01	6.23	4.29	.1732	.3977	.6373
AGE17	2	1	1.84	0.03	0.02	.1753	.8645	.8874
AGE17	3	2	7.30	4.44	3.05	.0260	.1089	.2172
AGE17	4	1	947.62	182.14	125.43	.0000	.0000	.0000
AGE17	5	2	17.95	10.35	7.13	.0001	.0056	.0283
AGE17	6	1	0.01	1.22	0.84	.9411	.2696	.3596
AGE17	7	2	3.13	3.87	2.66	.2088	.1445	.2639
AGE17	8	2	5.15	7.40	5.09	.0763	.0248	.0783
AGE17	9	17	11708.6	350.66	241.49	.0000	.0000	.0000
AGE17	10	1	10.61	5.34	3.68	.0011	.0209	.0552
AGE17	11	2	2.17	2.75	1.89	.3379	.2534	.3885
AGE17	12	2	276.08	181.79	125.19	.0000	.0000	.0000
AGE17	13	1	0.05	0.04	0.03	.8314	.8437	.8700
AGE17	14	9	129.18	43.21	29.76	.0000	.0000	.0005
AGE17	15	12	112.17	23.05	15.87	.0000	.0273	.1971
AGE17	16	12	1389.07	305.93	210.68	.0000	.0000	.0000
AGE17	17	6	9.01	6.23	4.29	.1732	.3977	.6373

Appendix C

Balanced Effect F-Tests

Table C-1. Balanced Effect F-Tests for 9-Year-Olds

		Design Based		Unweighted		Weighted	
	d.f.	F	Prob	F	Prob	F	Prob
	(30 denominator d.f.)				(2457 denominator d.f.)		
N0222A							
Sex	1	0.77	.39	0.98	.32	0.82	.37
Race	1	7.76	.01	12.66	.00	5.33	.02
PARED	3	4.84	.01	1.64	.18	2.76	.04
N0227A							
Sex	1	2.21	.15	1.05	.31	1.85	.17
Race	1	15.19	.00	22.27	.00	14.64	.00
PARED	3	39.76	.00	15.75	.00	17.94	.00
N0222A							
Sex	1	0.89	.35	1.10	.29	0.95	.33
TOC	2	0.66	.52	0.30	.74	0.54	.58
PARED	3	6.27	.00	2.64	.05	3.62	.01
N0227A							
Sex	1	1.95	.17	0.98	.32	1.64	.20
TOC	2	2.76	.08	2.89	.06	3.23	.04
PARED	3	56.72	.00	19.25	.00	20.81	.00
N0305C							
Sex	1	1.79	.19	0.85	.36	1.99	.16
Race	1	77.32	.00	126.34	.00	126.25	.00
PARED	3	10.60	.00	10.69	.00	14.83	.00
N0317A							
Sex	1	0.01	.91	0.06	.81	.04	.85
Race	1	4.88	.03	2.90	.09	2.18	.14
PARED	3	2.09	.12	4.00	.01	3.93	.01
N0323A							
Sex	1	1.47	.23	0.06	.81	1.62	.20
Race	1	16.41	.00	80.81	.00	61.15	.00
PARED	3	3.49	.03	3.93	.01	2.94	.03
N0305C							
Sex	1	2.21	.15	1.22	.27	2.67	.10
TOC	2	1.14	.33	6.97	.00	3.45	.03
PARED	3	12.20	.00	14.09	.00	16.87	.00
N0317A							
Sex	1	0.02	.89	0.05	.82	.05	.82
TOC	2	0.19	.82	1.55	.21	.83	.44
PARED	3	2.03	.13	4.27	.01	3.99	.01

(continued)

Table C-1. (continued)

		Design Based		Unweighted		Weighted	
	d.f.	F	Prob	F	Prob	F	Prob
	(30 denominator d.f.)				(2457 denominator d.f.)		
NO323A							
Sex	1	1.80	.19	0.15	.70	2.04	.15
TOC	2	0.07	.93	2.21	.11	0.31	.73
PARED	3	3.95	.02	6.15	.00	3.81	.01
Mean	(30 denominator d.f.)				(4895 denominator d.f.)		
Sex	1	2.91	.10	0.26	.61	2.09	.15
Race	1	80.95	.00	133.84	.00	103.89	.00
PARED	3	26.08	.00	23.75	.00	27.93	.00
Sex	1	2.81	.10	0.28	.59	2.02	.15
TOC	2	1.66	.21	4.25	.01	3.30	.04
PARED	3	27.98	.00	32.23	.00	34.12	.00

Table C-2. Balanced Effect F-Tests for 13-Year-Olds

		Design Based		Unweighted		Weighted	
	d.f.	F	Prob	F	Prob	F	Prob
T0105A	(30 denominator d.f.)				(2416 denominator d.f.)		
Sex	1	1.85	.18	2.91	.09	2.65	.10
Race	1	24.27	.00	65.93	.00	44.76	.00
PARED	3	54.98	.00	29.21	.00	35.58	.00
T0110A							
Sex	1	13.89	.00	11.95	.00	12.66	.00
Race	1	14.49	.00	60.12	.00	52.20	.00
PARED	3	7.10	.00	9.86	.00	9.30	.00
T0105A							
Sex	1	1.68	.21	2.13	.14	2.12	.15
TOC	2	2.87	.07	0.98	.37	8.23	.00
PARED	3	56.33	.00	42.81	.00	43.05	.00
T0110A							
Sex	1	19.01	.00	12.23	.00	13.63	.00
TOC	2	17.79	.00	14.46	.00	29.00	.00
PARED	3	11.74	.00	16.29	.00	14.09	.00
T0203A							
Sex	1	18.72	.00	26.60	.00	46.80	.00
Race	1	27.94	.00	76.78	.00	62.74	.00
PARED	3	19.81	.00	25.10	.00	27.72	.00
T0223A							
Sex	1	0.01	.92	0.82	.37	0.01	.92
Race	1	22.27	.00	79.67	.00	46.15	.00
PARED	3	7.95	.00	6.90	.00	9.28	.00
T0224A							
Sex	1	4.93	.03	7.77	.01	9.12	.00
Race	1	81.61	.00	87.00	.00	70.14	.00
PARED	3	24.04	.00	12.42	.00	14.02	.00
T0203A							
Sex	1	17.37	.00	22.75	.00	42.33	.00
TOC	2	0.05	.95	1.50	.22	0.33	.72
PARED	3	38.50	.00	41.09	.00	44.29	.00
T0223A							
Sex	1	0.01	.93	0.43	.51	0.01	.94
TOC	2	0.16	.85	2.89	.06	0.33	.72
PARED	3	16.66	.00	15.58	.00	17.32	.00

(continued)

Table C-2. (continued)

		Design Based		Unweighted		Weighted	
	d.f.	F	Prob	F	Prob	F	Prob
TO224A							
Sex	1	4.04	.05	6.45	.01	8.02	.00
TOC	2	1.02	.37	5.92	.00	2.10	.12
PARED	3	35.04	.00	25.80	.00	27.30	.00
Mean	(30 denominator d.f.)				(4849 denominator d.f.)		
Sex	1	13.87	.00	16.94	.00	22.17	.00
Race	1	59.41	.00	294.95	.00	224.55	.00
PARED	3	57.31	.00	60.97	.00	68.98	.00
Sex	1	14.70	.00	15.45	.00	21.41	.00
TOC	2	10.47	.00	11.80	.00	20.44	.00
PARED	3	81.71	.00	106.12	.00	108.98	.00

Table C-3. Balanced Effect F-Tests for 17-Year-Olds

		Design Based		Unweighted		Weighted	
	d.f.	F	Prob	F	Prob	F	Prob
SO108A	(30 denominator d.f.)				(2288 denominator d.f.)		
Sex	1	5.46	.03	6.07	.01	4.63	.03
Race	1	57.77	.00	111.52	.00	89.05	.00
PARED	3	12.11	.00	9.93	.00	10.66	.00
SO117A							
Sex	1	0.28	.60	0.10	.75	0.35	.56
Race	1	28.03	.00	86.78	.00	75.18	.00
PARED	3	12.30	.00	9.24	.00	8.30	.00
SO121A							
Sex	1	3.47	.07	2.39	.12	5.66	.02
Race	1	1.11	.30	0.95	.33	1.65	.20
PARED	3	1.24	.31	1.87	.13	1.55	.20
SO108A							
Sex	1	4.99	.03	5.25	.02	3.56	.06
TOC	2	1.12	.34	5.14	.01	3.09	.05
PARED	3	34.05	.00	27.61	.00	25.79	.00
SO117A							
Sex	1	0.38	.54	0.15	.70	0.40	.53
TOC	2	1.89	.17	1.84	.16	1.69	.19
PARED	3	17.01	.00	23.20	.00	18.86	.00
SO121A							
Sex	1	3.40	.08	2.25	.13	5.40	.02
TOC	2	2.09	.14	1.06	.35	1.60	.20
PARED	3	1.12	.36	1.89	.13	2.00	.11
SO206A							
Sex	1	17.93	.00	20.70	.00	19.28	.00
Race	1	76.02	.00	111.94	.00	120.39	.00
PARED	3	34.91	.00	25.50	.00	27.88	.00
SO225A							
Sex	1	30.22	.00	18.22	.00	23.10	.00
Race	1	121.03	.00	77.23	.00	86.55	.00
PARED	3	27.81	.00	17.50	.00	22.25	.00
SO206A							
Sex	1	22.23	.00	27.48	.00	23.39	.00
TOC	2	0.22	.81	0.57	.57	0.43	.65
PARED	3	32.75	.00	51.61	.00	55.71	.00

(continued)

Table C-3. (continued)

		Design Based		Unweighted		Weighted	
	d.f.	F	Prob	F	Prob	F	Prob
SO225A							
Sex	1	32.82	.00	23.59	.00	27.07	.00
TOC	2	0.23	.79	1.39	.25	0.45	.64
PARED	3	37.05	.00	34.75	.00	41.03	.00
Mean	(30 denominator d.f.)				(4562 denominator d.f.)		
Sex	1	23.40	.00	18.80	.00	19.91	.00
Race	1	76.56	.00	270.28	.00	297.57	.00
PARED	3	28.58	.00	31.23	.00	37.90	.00
Sex	1	26.86	.00	23.43	.00	20.41	.00
TOC	2	5.34	.01	0.71	.49	6.25	.00
PARED	3	32.90	.00	76.10	.00	84.79	.00

Appendix D

Contingency Tables of Wald Statistic Sample Design Based Tests Versus Alternative Tests Accepted and Rejected at the 5% Significance Level

Table D-1. Design Based Versus SRS Linear Model Tests of NAEP Items for the Race * Sex * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	25	1	26
	Reject	1	5	6
	Total	26	6	32

13-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	20	3	23
	Reject	4	13	17
	Total	24	16	40

17-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	28	2	30
	Reject	1	9	10
	Total	29	11	40

All Ages

		Design Based		
		Accept	Reject	Total
SRS	Accept	73	6	79
	Reject	6	27	33
	Total	79	33	112

Table D-2. Design Based Versus Adjusted Linear Model Tests of NAEP Items for the Race * Sex * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	26	1	27
	Reject	0	5	5
	Total	26	6	32

13-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	21	5	26
	Reject	3	11	14
	Total	24	16	40

17-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	28	2	30
	Reject	1	9	10
	Total	29	11	40

All Ages

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	75	8	83
	Reject	4	25	29
	Total	79	33	112

Table D-3. Design Based Versus SRS Contrast Tests of NAEP Items for the Race * Sex * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	18	3	21
	Reject	2	13	15
	Total	20	16	36

13-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	13	3	16
	Reject	1	28	29
	Total	14	31	45

17-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	15	2	17
	Reject	1	27	28
	Total	16	29	45

All Ages

		Design Based		
		Accept	Reject	Total
SRS	Accept	46	8	54
	Reject	4	68	72
	Total	50	76	126

Table D-4. Design Based Versus Adjusted Contrast Tests of NAEP Items for the Race * Sex * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	19	3	22
	Reject	1	13	14
	Total	20	16	36

13-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	13	4	17
	Reject	1	27	28
	Total	14	31	45

17-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	16	3	19
	Reject	0	26	26
	Total	16	29	45

All Ages

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	48	10	58
	Reject	2	66	68
	Total	50	76	126

Table D-5. Design Based Versus SRS Linear Model Tests of NAEP Items for the Sex * TOC * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	11	2	13
	Reject	1	2	3
	Total	12	4	16

13-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	16	7	23
	Reject	2	7	9
	Total	18	14	32

17-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	10	8	18
	Reject	2	12	14
	Total	12	20	32

All Ages

		Design Based		
		Accept	Reject	Total
SRS	Accept	37	17	54
	Reject	5	21	26
	Total	42	38	80

Table D-6. Design Based Versus Adjusted Linear Model Tests of NAEP Items for the SEX * TOC * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	12	3	15
	Reject	0	1	1
	Total	12	4	16

13-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	18	8	26
	Reject	0	6	6
	Total	18	14	32

17-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	10	9	19
	Reject	2	11	13
	Total	12	20	32

All Ages

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	40	20	60
	Reject	2	18	20
	Total	42	38	80

Table D-7. Design Based Versus SRS Contrast Tests of NAEP Items for the Sex * TOC * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	5	5	10
	Reject	0	8	8
	Total	5	13	18

13-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	9	2	11
	Reject	2	23	25
	Total	11	25	36

17-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	9	4	13
	Reject	1	22	23
	Total	10	26	36

All Ages

		Design Based		
		Accept	Reject	Total
SRS	Accept	23	11	34
	Reject	3	53	56
	Total	26	64	90

Table D-8. Design Based Versus Adjusted Contrast Tests of NAEP Items for the Sex * TOC * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	5	8	13
	Reject	0	5	5
	Total	5	13	18

13-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	11	7	18
	Reject	0	18	18
	Total	11	25	36

17-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	9	6	15
	Reject	1	20	21
	Total	10	26	36

All Ages

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	25	21	46
	Reject	1	43	44
	Total	26	64	90

Table D-9. Design Based Versus Adjusted Linear Model Tests for Mean Scores for the Race * Sex * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	3	3	6
	Reject	0	2	2
	Total	3	5	8

13-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	6	0	6
	Reject	0	2	2
	Total	6	2	8

17-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	5	1	6
	Reject	0	2	2
	Total	5	3	8

All Ages

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	14	4	18
	Reject	0	6	6
	Total	14	10	24

Table D-10. Design Based Versus SRS Linear Model Tests of Mean Scores for the Race * Sex * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	3	3	6
	Reject	0	2	2
	Total	3	5	8

13-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	6	0	6
	Reject	0	2	2
	Total	6	2	8

17-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	5	1	6
	Reject	0	2	2
	Total	5	3	8

All Ages

		Design Based		
		Accept	Reject	Total
SRS	Accept	14	4	18
	Reject	0	6	6
	Total	14	10	24

Table D-11. Design Based Versus SRS Contrast Tests of Mean Scores for the Race * Sex * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	2	0	2
	Reject	0	7	7
	Total	2	7	9

13-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	3	0	3
	Reject	0	6	6
	Total	3	6	9

17-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	2	0	2
	Reject	0	7	7
	Total	2	7	9

All Ages

		Design Based		
		Accept	Reject	Total
SRS	Accept	7	0	7
	Reject	0	20	20
	Total	7	20	27

Table D-12. Design Based Versus Adjusted Contrast Tests of Mean Scores for the Race * Sex * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	2	1	3
	Reject	0	6	6
	Total	2	7	9

13-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	3	0	3
	Reject	0	6	6
	Total	3	6	9

17-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	2	0	2
	Reject	0	7	7
	Total	2	7	9

All Ages

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	7	1	8
	Reject	0	19	19
	Total	7	20	27

Table D-13. Design Based Versus SRS Linear Model Tests of Mean Scores for the Sex * TOC * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	0	4	4
	Reject	0	4	4
	Total	0	8	8

13-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	3	1	4
	Reject	0	4	4
	Total	3	5	8

17-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	4	1	5
	Reject	1	2	3
	Total	5	3	8

All Ages

		Design Based		
		Accept	Reject	Total
SRS	Accept	7	6	13
	Reject	1	10	11
	Total	8	16	24

Table D-14. Design Based Versus Adjusted Linear Model Tests of Mean Scores for the Sex * TOC * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	0	7	7
	Reject	0	1	1
	Total	0	8	8

13-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	3	3	6
	Reject	0	2	2
	Total	3	5	8

17-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	5	1	6
	Reject	0	2	2
	Total	5	3	8

All Ages

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	8	11	19
	Reject	0	5	5
	Total	8	16	24

Table D-15. Design Based Versus SRS Contrast Tests of Mean Scores for the Sex * TOC * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	2	1	3
	Reject	0	6	6
	Total	2	7	9

13-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	2	0	2
	Reject	1	6	7
	Total	3	6	9

17-year-olds

		Design Based		
		Accept	Reject	Total
SRS	Accept	3	0	3
	Reject	0	6	6
	Total	3	6	9

All Ages

		Design Based		
		Accept	Reject	Total
SRS	Accept	7	1	8
	Reject	1	18	19
	Total	8	19	27

Table D-16. Design Based Versus Adjusted Contrast Tests of Mean Scores for the Sex * TOC * PARED Cross-Classification

9-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	2	3	5
	Reject	0	4	4
	Total	2	7	9

13-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	2	3	5
	Reject	1	3	4
	Total	3	6	9

17-year-olds

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	3	2	5
	Reject	0	4	4
	Total	3	6	9

All Ages

		Design Based		
		Accept	Reject	Total
Adjusted	Accept	7	8	15
	Reject	1	11	12
	Total	8	19	27

Appendix E

Contingency Tables of Balanced Effects Sample Design Based Tests Versus Alternative Tests Accepted and Rejected at the 5% Significance Level

Table E-1. Design Based Versus Alternative Tests for Balance Effects

9-year-olds

		Design Based		
		Accept	Reject	Total
Unweighted	Accept	14	2	16
	Reject	3	11	14
	Total	17	13	30

		Design Based		
		Accept	Reject	Total
Weighted	Accept	13	1	14
	Reject	4	12	16
	Total	17	13	30

13-year-olds

		Design Based		
		Accept	Reject	Total
Unweighted	Accept	7	0	7
	Reject	2	21	23
	Total	9	21	30

		Design Based		
		Accept	Reject	Total
Weighted	Accept	7	0	7
	Reject	2	21	23
	Total	9	21	30

17-year-olds

		Design Based		
		Accept	Reject	Total
Unweighted	Accept	11	0	11
	Reject	1	18	19
	Total	12	18	30

		Design Based		
		Accept	Reject	Total
Weighted	Accept	9	1	10
	Reject	3	17	20
	Total	12	18	30

Table E-1. (continued)

All Ages

		Design Based		
		Accept	Reject	Total
Unweighted	Accept	32	2	34
	Reject	6	50	56
	Total	38	52	90

		Design Based		
		Accept	Reject	Total
Weighted	Accept	29	2	31
	Reject	9	50	59
	Total	38	52	90